When Linguists do Set Theory

A Cautionary Tale

K. P. Hart

Faculty EEMCS
TU Delft

BLAST2022, 12 August, 2022 (18:00-19:00 UTC)

Background

Back in October 2019 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).
The discussion was about a paper that argued that all books past, present, and future already exist, as members of product sets of such alphabets.

This is not what I want to talk about ...

Proper classes?

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

> *Then, appealing to the reasoning of Langendoen and Postal (1984), 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.

I needed to know more . . .

The Vastness of Natural Languages

"Langendoen and Postal (1984)" refers to the book *The Vastness of Natural Languages*, which argues that the collection of sentences in a Natural Language is not a set but larger in magnitude than any set.

The book devotes a whole chapter [Chapter 3] to arguing
(quite vociferously I might add)
that there should be no size restrictions on sentences in Natural Languages.

Occam's razor gets mentioned a lot, not so much as a guiding principle but more like a law that must not be violated.
"... since Occam's razor forbids the incorporation of useless complications ..."

The Vastness of Natural Languages

The mathematical meat is in Chapter 4, *The Analogy with Cantor's Results*; it contains the proof that natural languages are megacollections (their term for what we would call proper classes).

That chapter is also available in a paper called *Sets and Sentences* by the same authors (advantage over the book: *shorter*, and easily available via Langendoen's website).

(The book can be 'borrowed' at `archive.org`.)

Rules for languages

The authors give descriptions of rules that show how to combine constituents into Co-ordinate compound constituents using connectives.

The nature of these is not important for the first part of this talk.
(They'll come back to haunt us later.)
Think of sets of sentences being combined into sentences using connectives, commas, (semi-)colons, and what not.

And of course there are no rules that limit the cardinality of the sets of constituents that can be used.

Existence

Central to the argument that a Natural Language is a proper class is an existence theorem.

Claim

Every set of constituents has a Co-ordinate compound constituent.

The authors give a ‘straightforward’ argument for this.

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 have at least one sentence, so we'll let this one slide.

As an aside: this sentence is representative of the pontificating style that the authors adopted.

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: from "there is an individual" to "there are individuals of all sorts".

The real mathematical error: the authors use the absence of size-restricting axioms to 'deduce' that there are arbitrarily large individuals.

(It seems that by leaving out axioms you can prove more ...)

The existence proof, 4

- "W can then, in particular, have exactly k such constituents."
  So the seemingly arbitrary W we started with has become quite specific.

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.

  (That's part of the definition.)

The existence proof, 6

Brace yourselves. Remember our arbitrary set U of constituents?

- "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."
  Right ...
  At the outset W and U were completely unrelated.
  And no, a bijection does not make sets equal.
- "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 the Claim?
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 compound constituent of U.

That ‘a’ has become a ‘the’. (We’ll get back to that in a moment.)

Closure Principle

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.

Closure Principle

The Closure Principle implies that Natural Languages form proper classes ("megacollections")

The NL Vastness Theorem

NLs are not sets (are megacollections).

The proof is just
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.

This explains the title, *The Analogy with Cantor's Results*, of the chapter: if the universe of sets were a set then it would contain its power set.

However

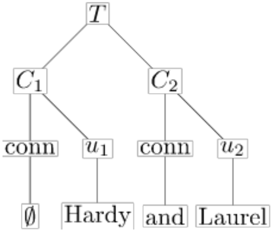
The proof of the existence claim uses, without justification, that there are compounds (sentences) of arbitrarily large cardinality.

We can at best treat that as an assumption, but that assumption is equivalent to the sentences forming a proper class, so . . .

the proof of the main result can be summarized as:
if the sentences in the natural language form a proper class
then the sentences in the natural language form a proper class

Coordinate compound constituents

Here is a picture of a *Co-ordinate compound constituent*



T is the Co-ordinate compound constituent;
C_1 and C_2 are conjucts;
'$\emptyset$' and 'and' are connectives (conjunctions);
'Hardy' and 'Laurel' are constituents, also called subconjuncts.

Projections and Projection Sets

Given a set U of constituents, of cardinality 2 or more, and a coordinate compound (constituent) T, all of some 'category' Q, we call T a *coordinate projection of U* and U the *projection set of T* if

1. each conjuct of T has an element of U as a subconjuct;
2. each element of U is a subconjuct of some conjuct of T;
3. no element of U appears more than once as a subconjunct of any conjunct of T; and
4. 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.

Projections and Projection Sets

According to the authors:
the last condition insures that different orders of the conjuncts are irrelevant.

Therefore coordinate projections of a set are unique *up to choice of the conjuction*, and one can speak of the coordinate projection of a set.

To some this may sound like a definition but it does not define anything.
What *is* the coordinate projection of $\{\mathrm{Smith}, \mathrm{Jones}\}$?

The authors simplify things by using only one conjunction, say 'and', so the coordinate projection of $\{\mathrm{Laurel}, \mathrm{Hardy}\}$ would be "And Hardy and Laurel".

Projections and Projection Sets

Some possibilities:

1. Use a global choice function to assign one coordinate projection to each set
2. Assign the whole set of coordinate projections to each set
3. The partially unordered set $\{\langle *, x\rangle : x \in U\}$, where $*$ is some connective, where (at least) x comes before y if x is shorter than y

The authors do not really define the coordinate projection.

And then there is the problem of 'length'.

What is a sequence?

At some point the authors introduce some order and talk about sequences.
Without ever defining them . . .

Here's an example

Many sentences

2 Jack$_1$ and his father$_2$ are visiting relatives.

3 Jack$_1$ and his father$_2$ and his father's father$_3$ are visiting relatives.

k Jack$_1$ and his father$_2$ and his father's father$_3$ and ... and his ... father$_k$ are visiting relatives.

∞ Jack$_1$ and his father$_2$ and his father's father$_3$ and ... and his ... father$_{\aleph_0}$ are visiting relatives.

Two things:

We can guess what sentence 4 will look like and we can specify a recursion that will produce sentence k for every natural number k, but ...
I wonder (actually strongly doubt) whether the authors can give an explicit description of sentence ∞.

The authors only talk about sentences of cardinal length: witness the $\aleph_0$.

A description L

The language L is decribed by

1. each sentence begins with the word a
2. each sentence end with the word b
3. In no sentence can the word a be immediately followed by the word a
4. In no sentence can the word b be immediately preceded by the string bb

A description

The first few sentences:

1. *ab*
2. *abb*
3. *abab*
4. *ababb*
5. *abbab*
6. *ababab*
7. *abbabb*
8. *abababb*
9. *ababbab*
10. *abbabab*

A recursive definition

Start with $B = \{ab, abb, abab, ababb, abbab\}$ (an inductive basis),
close off under the family F of transformations, where $F = \{A, B_1, B_2\}$ and:

1. $A : zbb \mapsto zbab$
2. $B_1 : zab \mapsto zabb$
3. $B_2 : zabab \mapsto zabbab$

That is how I would it, anyway.

The authors complicate things a bit.

Remarks about recursive definitions

The collection of all natural language sentences is called US.
"The family F of mappings that preserve all defining conditions on sentencehood in L includes the mappings A, B_1, and B_2 ..."
"Included in F are mappings other than those three, but these can be ignored ..."
An inductive collection over $\langle B, F \rangle$ is simply a collection that contains B and is closed under F.

"We now say that L is the intersection of all inductive collections over $\langle B, F \rangle$."

Back to our language

"The language L is the smallest collection of sentences drawn from US that is inductive over $\langle B, F \rangle$."

"If US is taken to be a finite set, then L is a finite subset of that set."

"If US is taken to be a denumerably infinite set, each member of which is of finite length. then L is the denumerably infinite set of sentences that is generated by the finite-state grammar $(*)$."

$$\begin{array}{ll} \text{a.} & S \rightarrow \{A, B\} b \\ \text{b.} & A \rightarrow (S) a \\ \text{c.} & B \rightarrow A\, b \end{array} \qquad (*)$$

Now, we (mathematicians) can prove that in this particular case, L is countably infinite and consists of finite sentences only.
However . . .

Uncountably many sentences?

"If sentences of L are allowed to be of length $\aleph_0$, it is easy to show that L has as many as $\aleph_1$ sentences and that there are sentences of length $\aleph_0$ that satisfy the defining conditions on sentencehood in L."

"We show this by establishing a one-to-one correspondence between a subset of L and the set of real numbers in the interval $0 \leqslant x < 1$."

Uncountably many sentences?

Here is the argument, verbatim:

> *Let the string ababb be interpreted as the numeral* 0 *in the binary expansion of a real number in the interval* $0 \leqslant x < 1$, *and let abbab be interpreted as the numeral* 1, *and suppose that a decimal point is understood before the leftmost a in any sentence of L.*
>
> *Thus the string ababb* $= .0 = 0$, *abbab* $= .1$, *ababbababb* $= .00 = 0$, *ababbabbab* $= .01 = 0$, *abbababab b* $= .10 = .1$, *abbabaabbb* $= .11$, *etc.*
>
> *Since, by assumption, sentences of L can be of length* $\aleph_0$, *there is a distinct sentence of L for each real number in the interval* $0 \leqslant x < 1$.

To be sure: the authors do not exhibit a single infinitely long sentence in *L*, nor do they seem to realize that having to have a *b* at the end makes it impossible to create infinite binary expansions as we know them.

There is more . . .

The authors leave ‘sequence’ essentially undefined; a whole chapter on transfinite sequences gives me the impression that ‘sequence’ is somehow synonymous with ‘linearly ordered set’ but that even sets like $\mathbb{Q}$ and $\mathbb{Z}$ have ‘length’ $\aleph_0$.

So length is not order type, but cardinality.

But I think I filled my time.

Light reading

Website: `https://fa.ewi.tudelft.nl/~hart`

K. P. Hart,
A critique of 'The vastness of Natural Languages', Lingbuzz 006052, June 2021

D. Terence Langendoen and Paul M. Postal,
The Vastness of Natural Languages, (1984) Oxford: Basil Blackwell.

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

Paul M. Postal,
Books, Lingbuzz 004733, August 2019.