How big are natural languages?

CityU/CTL Colloquium

7 November 2008

Terry Langendoen
Visiting Professor, CityU/CTL &
Professor Emeritus of Linguistics, U Arizona

Natural languages are infinite in size

- “Infinity is one of the most fundamental properties of human languages, maybe the most fundamental one.”
  - Lasnik, Howard. 2000. *Syntactic structures revisited: Contemporary lectures on classic transformational theory*. Cambridge, MA: MIT Press. Quotation is from p. 3.
- “FLN [Faculty of language–narrow sense] takes a finite set of elements and yields a potentially infinite array of discrete expressions…. The core property of discrete infinity is intuitively familiar to every language user.”
  - Hauser, Marc, Noam Chomsky & Tecumseh Fitch. 2002. The faculty of language: What is it, who has it, and how did it evolve? *Science* 298: 1569–1579. Quotation is from p. 1571.

Or are they?

- “Contrary to popular belief, it has never been shown that natural languages have infinitely many expressions.”
  - Pullum, Geoffrey and Barbara Scholz. 2005. Contrasting applications of logic in natural language syntactic description. In Petr Hájek, Luis Valdés-Villanueva & Dag Westerthål (eds.), *Logic, methodology and philosophy of science 2003*, 481-503. London: KCL Publications. Quotation is from p. 495.

Expressions of the ‘popular belief’ in natural language infinity

- Pullum & Scholz offer a sampler of presentations of the argument for the belief in natural language infinity (NLI) over the past 35 years, quoting such authorities as:

Emmon Bach	Edward Stabler
Steven Pinker	Andrew Carnie

What do Pullum & Scholz find wrong with these arguments?

- Pullum & Scholz propose that these justifications for NLI are all variants of an argument form they call the ‘Master Argument for language infinity’, and that it either unsound or circular.
- Here’s the Master Argument (omitting a few nonessential words and replacing reference to English with “any natural language”)….

The ‘Master Argument for language infinity’

- There is at least one well-formed expression in any natural language that has size greater than zero.
- For all *n*, if some well-formed expression in that language has size *n*, then [another] well-formed expression has size greater than *n.*
- Therefore, for every *n* there are well-formed expressions with size greater than *n* (i.e., the set of well-formed expressions in that language is countably infinite).
  - Pullum & Scholz (2005: 496), with changes as noted on the previous slide.

Why does the Master Argument fail?

- It assumes that there is a set of expressions in the language under analysis. That set must be assigned some size.
  - If it is finite, then the argument fails because it is unsound.
  - If it is infinite, then the argument begs the question.

What about a reductio ad absurdum *argument instead*?

- Assume that some natural language has only finitely many expressions, each finitely long.
- Then there is some number *n* that is the length of the longest such expression (or expressions if there is more than one).
- However, an expression of length greater than *n* in that language can be constructed from one of those expressions by a rule or principle of that language.
  - E.g. Hauser, Chomsky & Fitch (2002: 1571) propose embedding such an expression in the position of the *x* in *Mary thinks that x*.
- This is a contradiction.
- Therefore the assumption that the language has only finitely many finitely long expressions must be rejected. It must have infinitely many ones.

So who's right, or is the jury still out?

- My response is that the jury is still out.
- By themselves, the existence of operations that yield new and longer expressions in a language does not show that the language is infinite.
- It must also be shown that the language is ***closed*** under those operations. Showing that the grammatical description is closed is not sufficient.

Closure is part of an ***inductive definition*** *of a set*

a) Base case: $S_0 \subseteq S$

b) Recursive step: for all x, if $x \in S$ then $\Phi x \in S$

c) Closure: if T also satisfies a) and b), then $S \subset T$; i.e., S is the smallest set to satisfy a) and b)

Casting the Hauser et al. example as an inductive definition to show that English is infinite

a) Base case: {*Amy is a doctor*} $\subseteq D$

b) Recursive step: for all x, if $x \in D$ then *Mary thinks that x* $\in D$

c) Closure: if E also satisfies a) and b), then $D \subset E$; i.e., D is the smallest set to satisfy a) and b)

- Therefore, D = (*Mary thinks that*)* *Amy is a doctor*, which is an enumerably (or countably) infinite set.
- We may take *E to* be the set of expressions of English, since that set also satisfies a) and b). Then, according to c) $D \subset E$, so that E is also an infinite set.

Pseudocode for producing the inductively defined set S

```
begin
m := 0
input n
** Require that n be non-negative integer. **
do while m < n
   say "Mary thinks that "
   m := m+1
end do
say "Amy is happy."
end
```

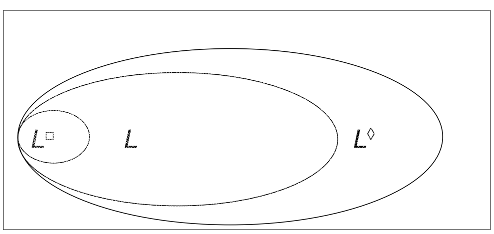
A reason to be skeptical of inductive definitions of natural languages

- It is reasonable to assume that there is a bound on the size of *n* for any device that carries out an implementation such as the one shown on the previous slide for producing or comprehending all of an enumerably infinite set of expressions.
- But is that bound a matter of linguistic **competence** or **performance**?

Redescribing the situation

- For every natural language *L*, there is a finite set of expressions ***known*** to belong to it.
  - Call that set $L^{\square}$, the ***necessary*** members of *L*.
- A grammatical description ***projects*** from $L^{\square}$ a possibly infinite set of ***possible*** members of *L*.
  - Call that set $L^{\Diamond}$, the possible members of *L*, and assume for purposes of discussion that the description is correct, at least insofar as it is based on $L^{\square}$ and makes reasonable assumptions.
- The language *L* itself falls somewhere in between $L^{\square}$ and $L^{\Diamond}$, presumably bigger than $L^{\square}$ and possibly smaller than $L^{\Diamond}$.

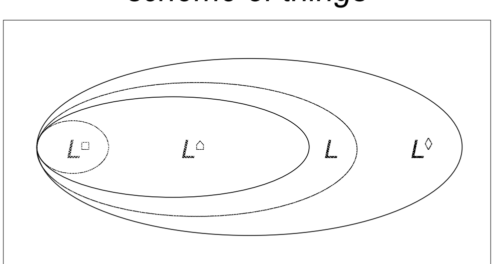
How far can we project membership in L beyond $L^{\square}$?



What do people need from the language they speak?

- The set of expressions that contains everything that people need from their language *L* may be presumed to be part of that language, i.e. is contained in *L*. Label that set $L^{⌂}$.
- Presumably, $L^{⌂}$ is bigger than $L^{\square}$.
- If $L^{⌂}$ is infinite, then so is *L*, establishing NLI.
- If $L^{⌂}$ is finite, then no conclusion can be drawn about NLI, but doubt would be cast on it.

Positioning $L^{⌂}$ in the scheme of things



Sapir's answer to the question of need

- Sapir contended that every natural language provides "a complete system of reference" for human experience, so that for any of its speakers, "no matter how original or bizarre his idea or fancy, the language is prepared to do his work".
  - Sapir, Edward. 1949 [1924]. The grammarian and his language. *American Mercury* 1: 149–155. Reprinted in David Mandelbaum (ed.), *Selected writings of Edward Sapir in language, culture and personality*, 150–159. Berkeley: University of California Press. Quoted snippets are from p. 153 of the reprint.

"The outstanding fact about any language is its formal completeness."

- Accordingly, "The world of linguistic forms, held within the framework of a given language, is a **complete system of reference**, very much as a number system is a complete system of quantitative reference…."
  - Both quotations from Sapir (1949 [1924]: 153). (emphasis mine)

How can Sapir's formal completeness criterion be understood?

- I suggest closure under an inductive definition, on analogy with such a definition for the natural numbers so as to provide a "complete system of reference" for quantity.
  a) Base case: $1 \in Z$
  b) Inductive step: if $n \in Z$, then $n+1 \in Z$
  c) Closure: Z is the smallest set to satisfy a) and b).

What this means for NLI

- My assumption that the expressions in $L^{⌂}$ are closed under certain operations needed for expressive power may minimally satisfy Pullum & Scholz's (2005: 496) requirement that the claim that a natural language "actually contains all the members of the closure of some set of … expressions under certain lengthening operations."

Did Sapir have anything to say about the size of natural languages?

- Not directly as far as I have been able to determine, but he did claim that:
  - Distinct "functionally equivalent expressions" for the same concept (e.g. *it's fun to laugh*) can be continued "***ad infinitum***", and
  - "All languages are set to do all the symbolic and expressive work that language is good for, **either actually or potentially**."
    - Sapir (1949 [1924]: 155), emphases mine.

Hauser et al. on the relation between natural language and natural numbers

- Hauser, Chomsky & Fitch also consider the analogy between natural language and natural numbers but from an evolutionary perspective.
  - They contend that "the innovation that yielded the faculty of language was the evolution of the **computational system** that links the [human sensory-motor and conceptual-intentional interfaces]".
    - Hauser, Chomsky & Fitch (2002: 1578), emphasis mine.

Requirements the computational system according to Hauser et al.

- "The computational system must [be able to]
  i. construct an infinite array of internal expressions from the finite resources of the conceptual-intentional system, and
  ii. provide the means to externalize them at the sensory-motor end."
    - Hauser, Chomsky & Fitch (2002: 1578)

Comparison of Sapir and Hauser et al. on NLI

- Sapir maintains that natural languages are formally complete, like arithmetic and geometric systems of reference, but does not explicitly conclude that they are thereby infinite.
- Hauser, Chomsky & Fitch take as given that natural languages are enumerably infinite, and suggest that this property may have resulted from an evolutionary “innovation” that developed “to solve other computational problems such as navigation, number quantification, or social relationships”.
  – Hauser, Chomsky & Fitch (2002: 1578)

I’ll stay with Sapir from here on

- Since our goal is to determine whether NLI is correct, we cannot assume that it is, on pain of begging the question, as in the Master Argument for language infinity.
- Therefore I put aside further consideration of Hauser, Chomsky & Fitch’s position, despite its interest.

The formal completeness criterion by itself does not settle the question of NLI

- A formally complete set can be:
  – Finite
    • The set of any given person’s ancestors
    ➢ The set M_0 of truth tables (‘distinct meanings’) of propositional logic PL_n with finitely many atomic propositions $P_f = \{P_0, \ldots, P_n\}$.
  – Enumerably infinite
    • The set of natural numbers
    ➢ The set of well-formed formulas (‘expressions’) of PL_n according to its standard syntax
    ➔ The set M_1 of finitely-sized truth tables of propositional logic PL_∞ with infinitely many atomic propositions $P_\infty = \{P_0, \ldots\}$.
  – Non-enumerably (transfinitely) infinite
    • The set of real numbers
    ➔ The set M_2 of all truth tables of PL_∞.

How many paraphrases for a single meaning does a language need?

- This question arises in case we decide that $L^{\triangle}$ has at least the expressive resources of PL_n given the distinction between the items bulleted “➢” in the preceding slide.
- Two possible answers:
  – It’s sufficient to have only a finite number of expressions for every distinct meaning.
    • $L^{\triangle}$ is still finite, so the question of NLI remains open.
  – It’s necessary to have an unlimited number of ways of expressing the same thing.
    • $L^{\triangle}$ is infinite, so the question of NLI is decided in the affirmative.

The possibility arises that natural languages are non-enumerably infinite

- Suppose that $L^{\triangle}$ has at least the expressive resources of PL_∞, as in the items bulleted “➔” two slides back; i.e. grant that languages are at least enumerably infinite in size.
- The set M_1 is closed under logical negation, conjunction and disjunction but such that at most finitely many members of P_∞ occur in each member of M_1. M_1, like P_∞, is enumerably infinite.
- The set M_2 is closed under those operations without restriction. M_2 is non-enumerably infinite.

Possible utility for expressions involving unbounded conjunction

- Expressions of mutual belief
  – A believes that *p*, and B believes that *p*, and A believes that B believes that *p*, and B believes that A believes that *p*, and A believes that B believes that A believes that *p*, and …
    • Adapted from Pullum & Scholz (2005: 497), citing Schiffer (1972) and Joshi (1982).