7. Just how big are natural languages?

D. Terence Langendoen

1. The question of natural language infinity

The assumption that natural languages are comprised of an infinite set of expressions is widely held, and viewed as a characteristic that must be accounted for by any theory of natural language.[1] Pullum & Scholz (2005, 2009) have recently argued, however, that no adequate demonstration for natural language infinity (NLI) has ever been made and that the cardinality of the set of all the expressions in a language is not important for the formulation of grammars for natural languages (2005: 497). In this section, I examine their argument against the claims for NLI and conclude that they are correct in asserting that the question of NLI remains open, though perhaps not exactly for the reasons they provide. However I do not agree with them about the lack of importance of the question of NLI, and in the remainder of the paper attempt to show how it might be answered.

Pullum & Scholz (2005: 495) threw down the gauntlet when they asserted that "[c]ontrary to popular belief, it has never been shown that natural languages have infinitely many expressions". They contend that what they call the Master Argument for language infinity is the basis for all the putative demonstrations of NLI, and that it fails because it is either unsound or begs the question. They also note that their point is not new, but a paraphrase of an argument that Paul Postal and I published some 25 years ago (Langendoen & Postal 1984: 30–35). Since Paul and I went on to make an even stronger claim about how many expressions a natural language has, namely that it has transfinitely (i.e. more than denumerably infinitely) many, I find myself in the somewhat odd position of having

[1] I thank three anonymous reviewers, who convinced me that the original version of this paper needed a major overhaul. This material is based in part upon work that was supported while I was serving at the National Science Foundation. Any opinion and conclusions are those of the author and do not necessarily reflect the views of the National Science Foundation.

argued that NLI is correct, but also of having helped formulate an argument that purports to deny that anyone has ever shown that it is correct.

Pullum & Scholz (2009) elaborate on this point by showing that NLI cannot be established any of a variety of arguments for infinite size, such as inductive generalization or mathematical induction. Similarly, it cannot be established by a *reductio ad absurdum* of the form (1)–(3):

(1) Assume that natural languages have at most finitely many finitely large expressions;
(2) Show that the assumption together with other known properties of natural languages leads to a contradiction, namely that all expressions have an upper bound on size, but that there is at least one larger one; and
(3) Conclude that natural languages must have infinitely many expressions.

The success of the *reductio* argument depends on identifying the correct known properties. As Pullum & Scholz (2005, 2009) point out, the known property that is typically appealed to in arguments for NLI is the existence of operations that iteratively increase the size of expressions and that in doing so preserve well-formedness. Certainly, as they also point out, if a language is *closed* under one or more of those operations, so that they are genuinely recursive, not just iterative, it is infinite. However from the fact that one's grammatical model is closed under such an operation, it does not follow that the language it models is. Without a demonstration of closure under a particular operation *for the language itself,* the issue of the correctness of NLI remains open. In formulating our argument for NLI, Postal and I asserted that English and other natural languages are closed under two iterative size-increasing operations; however, whether we were successful in demonstrating it is also open to question. I return to that issue in Section 3. In the next section, I consider a way of determining whether a natural language is infinite and conclude that it is possible for some natural languages to be infinite while others aren't.

2. Determining the size of natural languages

Determining the size of a natural language requires projecting beyond what is known about it from the direct evidence we have at hand. Our direct evidence comes in the form of judgments about particular expressions: whether they are grammatical in that language, whether they have such-and-such conditions of use in that language, etc. For any natural language L, let $L^{\square}$ represent the finite set of expressions known to belong to L

on the basis of such direct evidence. Given that $L^{\square}$ provides indirect evidence for genuinely recursive size-increasing operations, standard generative models project a denumerably (countably) infinite set $L^{\Diamond}$ of 'possible' members of L. By not distinguishing the models from the language, proponents of such models conclude that $L = L^{\Diamond}$ and so is infinite. As noted above, that conclusion may be correct, but an argument is still needed to show that the models do not overgenerate. In the absence of such an argument, all we can conclude is that L lies somewhere between $L^{\square}$ and $L^{\Diamond}$, and so may be either finite or infinite. The question we now face is this:

Q1. How far can we project membership in L beyond $L^{\square}$ with a reasonable degree of certainty?

One promising approach is to attempt to identify on the basis of $L^{\square}$ those expressions $L^{\hat{}}$ that are needed by speakers of L, since it is reasonable to assume that L contains every expression that its speakers will ever have occasion to need. Taking this approach, we then ask:

Q2. What do people need from the languages they speak?

It is clear that a large finite set of expressions will suffice to satisfy the expressive needs of anyone with finite temporal and physical resources, i.e. everyone. However it is appropriate to abstract away from those resource limitations to ask what people might need if those limitations were removed. The simplest and most striking answer to this question was given by Sapir (1949 [1924]: 153), who contended that every natural language has the property of "formal completeness", thereby providing "a complete system of reference" for human experience, on analogy with numerical and geometric systems of reference for quantity and space, 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".[2]

---

[2] Von Fintel & Matthewson (2008: 142–146) consider Sapir's'formal completeness' thesis to be a forerunner of Katz's (1976) 'effability' thesis for natural languages — that every language is capable of expressing every meaning. If they are right about this, it would not be inappropriate to refer to the effability thesis as the 'Sapir-Katz hypothesis'. However Sapir's thesis explicitly relates a language's expressive power to speaker's need in a way that Katz's does not. On the other hand, Katz considers effability the defining characteristic of natural languages, whereas Sapir simply considers formal completeness "[t]he outstanding fact" about them.

Although Sapir's answer is a simple one, it is not immediately clear how to apply it to the problem under consideration, except perhaps by exploiting his mathematical analogy, which he says "is by no means as fanciful as it appears to be". Since the analogous systems are all complete in the sense that they are closed under the relevant operators, e.g. addition for arithmetic, it seems reasonable to construe formal completeness to mean that linguistic systems are comparably closed under the relevant operators.[3] If we accept Sapir's formal completeness hypothesis as just interpreted as the basis for identifying membership in $L^{\hat{}}$, our question next becomes:

Q3. What is the size of the set of expressions that is closed under the relevant operators?

Having noted above that closure under iteratively size-increasing operations results in an infinite set, it would appear that the answer is just that. For example, from the occurrence of tautocategorial embedding — Pullum & Scholz's (2005) example of an iterative size-increasing operation that preserves well-formedness in English — in members of $L^{\square}$ together with the judgment that the operation is needed for L to do the expressive work for its speakers, $L^{\hat{}}$ is closed under that operation, and is thereby infinite. However I also interpret Sapir's hypothesis as consistent with the view, argued for by Everett (2005), that the expressive needs of one linguistic community can differ from that of another, since the languages that meet those different needs can nevertheless all be formally complete. If for example the community speaking language P does not need the ability to refer to distant ancestors, whereas the one speaking language Q does, $P^{\hat{}}$ may still be formally complete in Sapir's sense, being analogous, say, to an arithmetic system over the set of positive integers, with $Q^{\hat{}}$ analogous to one over the full set of integers. One can even imagine the community speaking P to have such limited expressive needs that $P^{\hat{}}$, though formally complete, is finite, analogous to an arithmetic system over the set of positive integers modulo some large, but finite number.

For example, imagine that speakers of P need no more expressive resources than a (possibly large, but finite) set of k simple affirmative sentences and the operations of negation, conjunction and disjunction defined

---

[3] Starting from the assumption of NLI, Hauser, Chomsky & Fitch (2002: 1571) consider natural language to be "directly analogous to the natural numbers" but nothing like formal completeness or closure figures in their account of human linguistic capacity.

over them. Then the smallest language $P^{\hat{}}$ that is closed under these operations is much larger, but still finite, having 2^{2^k} members, in which there is exactly one expression for each logically distinct member of $P^{\hat{}}$. Of course *P* may be larger, allowing for paraphrase, but the method proposed here for determining whether *P* is infinite would not lead to a definite conclusion one way or the other.

3. Can natural languages be bigger than denumerably infinite?

The starting point for Postal's and my argument for NLI is not much different from the hypothetical example in the previous paragraph. Instead of a finite set of simple affirmative sentences as the base of operations for *P*, it proposes an infinite set P_0 based on closure over a single iterative tautocategorial embedding operation. Granted that this starting point begs the question of establishing NLI, we had a different concern, namely to show that natural languages are not merely denumerably infinite but transfinite in size. The only operation we considered over this base was conjunction, and we asserted that *P* is closed under that operation, by which we meant the condition in (4) (an update of the starting point for our Closure Principle for Coordinate Compounding), in which the absence of a final member of the list of expressions in the antecedent is critical; it indicates that there is no finite bound on the number of conjuncts in members of *P*. From (4) it follows that *P* has nondenumerably many expressions.

(4) If p_1, p_2, ... are in P_0, then their conjunction is in *P*.

Our argument for (4) was based not on consideration of expressive need, but rather on economy of description: The simplest empirically adequate grammatical account of conjunction does not limit the number of unconjoined expressions that can be conjoined, so (4) is to be preferred to any account that does limit it, for example to a finite number of conjuncts.[4] Moreover, if we think of conjunction in *P* as logical rather than grammatical, then (4) can be recast as a valid entailment schema as in (5), where p_1, p_2, ... is a possibly infinite sequent.

---

[4] Conjoined expressions can also be members of a conjunction, but not recursively so (Langendoen 1998), so for convenience they are left out of the formulation in (4)., If the antecedent list in (4) is construed as the members of a set, it also does not provide for conjuncts to be repeated.

(5) $p_1, p_2, \ldots \models p_1 \& p_2 \ldots$

However despite its elegance, our argument has not convinced many linguists of its correctness, even among those who uncritically accept NLI.[5] There are, I believe, two reasons for this. One is the entrenched dogma, as Pullum & Scholz (2005) put it, that every expression in a natural language is a finitely-sized object; see also Dale (1996) and Hintzen & Uriagereka (2006). The other is that no convincing need has been identified for infinitely-sized expressions. Pullum & Scholz (2005: 497) make the best case that I am aware of for the usefulness of infinitely-sized expressions, namely for characterizing the notion of mutual belief, citing the work of Schiffer (1972) and Joshi (1982).[6]

Postal and I did not stop at (4) in our formulation of the principle of closure under conjunction, but went on to propose that for every nonempty, nonsingleton set of expressions of a natural language, it contains a conjunctive expression having every member of that set as a conjunct, from which it follows that natural languages are proper classes, making them too large to be considered sets. However in the absence of a need for expressions of a size greater than that of the denumerably-infinitely-long conjunctions characterized by (4), I conclude that they are not part of $L^{\cap}$ for any natural language L.[7]

4. Conclusion

Determining the size of a natural language is not as easy as simply declaring that there is no longest expression in any language and saying as a

---

[5] Though it has found some resonance in computer science; see for example Zeitman (1993).

[6] Uriagereka (2005) considers, but does not formalize, the possibility that the attachment of disjuncts, the class of adjuncts that do not scope over one another, gives rise to infinitely-sized expressions, and perhaps more interestingly, to infinitely large forms of interpretation expressible with finite phonologies. Even if all that is correct, it still remains to be seen whether a need for them can be identified.

[7] Pullum & Scholz (2005: 498, n. 15) give as a reason for not characterizing natural languages as proper classes the fact that the closure principle that leads to that result is unstatable as a Model-Theoretic Syntax constraint. If I read them correctly, the weaker closure principle in (4) is also unstatable, but there is another way in that framework for admitting denumerably infinitely-sized expressions.

result “Clearly it’s denumerably infinite”, and we may be grateful to Pullum & Scholz (2005, 2009) for pointing that out. The greatest difficulty is in finding and agreeing upon a basis for determining whether a language is closed under one or more of its iterative size-increasing operations, and if so how. If the basis for doing so is whether the results of the operations are needed by the community of speakers of that language, as I have suggested, following Sapir, there is still room for dispute about which operations should be counted, and about what conclusion to draw if it should turn out that there are no such operations in a particular language. The question of whether the sets of expressions of particular natural languages are finite, denumerably infinite or nondenumerably infinite (of the cardinality of the real numbers) remains open.

Furthermore, given that the question of the size of natural languages remains a matter of dispute, we need to look more deeply at the relation between natural languages and mathematical systems than simply the parallel between the enumeration of their members (expressions on the one hand and integers on the other, for example), as Sapir did when he developed the notion that natural languages are formally complete.

References

Dale, Russell Eliot.
1996 The theory of meaning. Ph. D. diss., Philosophy Program, Graduate Center of the City University of New York. https://webspace.utexas.edu/deverj/personal/test/theoryofmeaning.pdf (accessed 2007-04-01).

Everett, Daniel L.
2005 Culture constraints on grammar and cognition in Pirahã: Another look at design features of human language. *Current Anthropology* 46: 621–646.

Hauser, Marc, Noam Chomsky and W. Tecumseh Fitch
2002 The faculty of language: What is it, who has it, and how did it evolve? *Science* 298: 1569–1579.

Hintzen, Wolfram and Juan Uriagereka
2006 On the metaphysics of linguistics. *Erkenntnis* 65: 71–96.

Joshi, Aravind
1982 Mutual beliefs in question-answer systems. In Neil Smith (ed.), *Mutual Knowledge*, 181–197. London: Academic Press.

Katz, Jerrold
1976 A hypothesis about the uniqueness of natural languages. In: Steven Harnad, Horst Steklis and Jane Lancaster (eds.), *Origins*

and Evolution of Language and Speech = Annals of the New York Academy of Sciences 280: 33–41.

Langendoen, D. Terence
1998 Limitations on embedding in coordinate structures. *Journal of Psycholinguistic Research* 27: 235–259.

Langendoen, D. Terence and Paul Postal
1984. *The Vastness of Natural Languages*. Oxford: Blackwell.

Pullum, Geoffrey and Barbara Scholz
2005 Contrasting applications of logic in natural language syntactic description. In: Petr Hájek, Luis Valdés-Villanueva and Dag Westersthål (eds.), *Logic, Methodology and Philosophy of Science 2003: Proceedings of the 12th International Congress*, 481–503. London: KCL Publications.

Pullum, Geoffrey and Barbara Scholz
2009 Recursion and the infinitude claim. This volume.

Sapir, Edward.
1949 Reprint. The grammarian and his language. In: David Mandelbaum (ed.), *Selected Writings of Edward Sapir in Language, Culture and Personality*, 150–159. Berkeley: University of California Press. Original edition, *American Mercury* 1: 149–155, 1924.

Schiffer, Steven
1972 *Meaning*. Oxford: Clarendon Press.

Uriagereka, Juan
2005 Adjunct space? Paper presented at the Prospects for dualism conference, Amsterdam.

von Fintel, Kai and Lisa Matthewson
2008 Universals in semantics. *The Linguistics Review* 25: 139–201.

Zeitman, Suzanne
1993 Somewhat finite approaches to infinite sentences. *Annals of Mathematics and Artificial Intelligence* 8(1–2): 27–36.