

Just how big are natural languages?

D. Terence Langendoen

Linguistics Program, National Science Foundation, and
Department of Linguistics, University of Arizona

The size of a natural language is not determined simply by the grammar of that language. This is true not just for the class of model-theoretic syntactic frameworks described in a recent series of papers by Pullum and Scholz, but also the class of what they call generative-enumerative syntactic frameworks, and which I refer to as production systems. Using the notion of inductive definition to characterize a natural language as a subset of a universal set of expressions, I show that the size of a language, whether finite, denumerably infinite, or nondenumerably infinite is determined by both the nature of the inductive definition and the size of the universal set.

Recursion defined inductively

If a human language is recursive in the intuitive sense of being a set of expressions made up of other expressions that also belong to that language without apparent limit, then it can be characterized by a set of *inductive definitions*.¹ Each inductive definition, which is a procedure for defining subsets L of a universal set, has three parts as in (I); cf. Zalabardo (2002: 41).

- (I) a. a *base* stipulating that certain members of U are also members of some L;
b. a series of *inductive clauses* each stipulating that the image of a function over the members of some L is also a member of some L, provided that that image is a member of U;
c. a *closure clause* stipulating each L is the smallest subset of U that satisfies (Ia) and (Ib); i.e., that no proper subset of each L satisfies (Ia) and (Ib), alternatively that each L is a proper subset of any other set that satisfies (Ia) and (Ib).

If U is denumerably infinite, then any subset of U defined inductively is also denumerably infinite, provided that certain conditions are met.² For example, let U be the denumerably infinite set of strings W^* over a set W of words containing at least the words *her*, *she*, *someone*, *to*, *visit*, *want* and *wants*. Then (I1) below inductively defines the denumerably infinite regular language $L1 = she\ wants\ (someone\ to\ want)^+ someone\ to\ visit\ her$. The production system (or generative grammar) (G1) is equivalent to (I1), by interpreting the last production listed as the base and the first and second as inductive clauses, closure being part of the definition of a production system.³

- (I1) a. *someone to visit her* \in L1a
b. i. $\delta(x) \in$ L1a for every $x \in$ L1a, where $\delta(x) = someone\ to\ want\ x$
ii. $\eta(x) \in$ L1 for every $x \in$ L1a, where $\eta(x) = she\ wants\ x$
c. i. if $K \subseteq W^*$ also satisfies (I1a) and (I1bi), then $L1a \subset K$
ii. if $K \subseteq W^*$ also satisfies (I1bii), then $L1 \subset K$

¹ A version of this paper has been submitted for publication. The material in it is based in part upon work supported while the author 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.

² In particular, the image of at least one of the functions is not in its range for all but finitely many applications. This condition is assumed throughout this discussion.

³ In (G1), N = nonterminal vocabulary (categories), T = terminal vocabulary (words), A = axioms (start symbols), P = productions (rules), as in Hopcroft and Ullman (1979).

(G1) $N = \{S, S'\}$

$T = \{her, she, someone, to, visit, want, wants\}$

$A = \{S\}$

$P = \{S \rightarrow she\ wants\ S', S' \rightarrow someone\ to\ want\ S', S' \rightarrow someone\ to\ visit\ her\}$

Assuming that English is a subset of W^* , satisfies (I1a) and (I1b), and contains some member of W^* that satisfies neither (I1a) nor (I1b), then by (I1c), $L1 \subset \text{English}$, and consequently English is also denumerably infinite.

Questioning denumerably infinite natural languages

Pullum and Scholz (2005: 16-17) contend that certain presentations of the argument that natural languages are denumerably infinite are circular. However the presentation based on (I1) is not circular, as the assumption that a natural language is a *subset* of W^* is consistent with its being finite. The only assumptions to which an objection can reasonably be raised against the argument from (I1) that English is denumerably infinite are given in (A).

(A) a. English satisfies (I1b).

b. The universal set of which English is a subset is denumerably infinite.

Pullum and Scholz also observe that one can grant “the existence of productive lengthening operations [inductive clauses such as those in (I1b)] in natural languages” (2005: 17), while denying that those languages are denumerably infinite. All one has to do is deny (Ab) for the languages in question.

One way of denying (Ab) is to assume that the universal set for natural languages is finite, for example the subset W^{3n+4} of W^* , containing all and only all the members of W^* of length $3n+4$ words or less, for some positive, finite n . Although, for example, there is an $x \in L1a$ for which $\delta(x) \notin L1a$, namely $(someone\ to\ want)^n\ someone\ to\ visit\ her$, its image under $\delta \notin W^{3n+4}$, so it does not constitute a counterexample to the inductive definition. If U is restricted to W^{3n+4} , I1 defines $L1$ as the finite set $\{she\ wants\ (someone\ to\ want)^m\ someone\ to\ visit\ her \mid 0 \leq m < n\}$, without any restriction having to be placed on the inductive clauses. Further, since English is assumed to be a subset of W^{3n+4} , it is also finite, despite being defined inductively. The same conclusion follows for a production system, what Pullum and Scholz call a *generative-enumerative syntax* (GES) system, as long as its productions are understood as parts of an inductive definition. The claim that a production system that incorporates an inductive clause necessarily defines a denumerably infinite language is false; whether it does depends on the choice of the universal set from which the members of the language it generates are drawn. If that set is denumerably infinite, then the language it generates is denumerably infinite, but if that set is finite, then the language it generates is finite.⁴

⁴ I specifically deny Pullum and Scholz’s contention that “IF WE ASSUME THAT A NATURAL LANGUAGE CAN ONLY BE CORRECTLY DESCRIBED BY A GES GRAMMAR, [emphasis in original] it immediately follows that the set generated by a grammar for English contains infinitely many expressions”. (2001: 35) Pullum and Scholz draw their conclusion from a reading of the GES literature, e.g., passages like: “We need to find a way of representing structure that allows for infinity.... Infinity is one of the most fundamental properties of human languages, maybe the most fundamental one.” (Lasnik 2000: 3) In that literature, the notions recursion and the denumerable infinity of the universal set are generally bound up together. One of the purposes of this paper is to separate these notions, so that the goal that Lasnik describes as the need “to find a way of representing structure that allows for infinity” is properly formulated. See also Langendoen and Postal (1984: 31-32), for a refutation of a claim similar to Pullum and Scholz’s.

Finally, Everett's (2007) assertion that Pirahã lacks recursion can be construed as a denial of (Aa), for any choice of inductive clause and with Pirahã replacing English.

Justifying denumerably infinite natural languages

Langendoen and Postal (1984: 30-42) surveyed the arguments for assuming (Ab), and concluded that the only convincing one was formulated by Katz (1966: 122), who observed that if a finite bound is placed on the size of the members of a natural language defined inductively, there would be structures, in fact infinitely many of them, with all the grammatical properties of the members of that language but which would be excluded from membership in that language simply because of their size. In my judgment, no effective counterargument to Katz's argument against discrimination on the basis of size has ever been given.

This of course is not to deny that recursion in natural languages is not constrained in the manner described by Joshi (2007). In fact, characterizing natural-language recursion inductively may help determine the nature of some of the constraints he discusses. For example, from the fact that both *who does she want someone to visit e* and *who does she want e to visit her* are well-formed in English (where *e* is an empty string bound by *who*), we can legitimately extend the base case for L1a to include both *someone to visit e* and *e to visit her*, but require the empty string to be bound in the application of the function η , so that, for example, *she wants someone to visit e* and *she wants e to visit someone* are not included in L1.⁵ Similarly, neither of the strings *who does she want e to want someone to visit e* nor *who does she want e to want e to visit her* belong to L1, if we assume that exactly one empty string can be bound by *who*.

Inductive definitions and nondenumerably infinite natural languages

Another way of denying (Ab) is to assume that W^* , rather than being too big, is too small, because there are linguistically motivated inductive clauses that define nondenumerably many members of a natural language. In this section I provide motivation for an inductive definition containing such a clause.

First, consider (I2), which is an inductive definition of a denumerably infinite subset L2 of W^* which is assumed also to contain the word *and*, and (G2), which is the production system corresponding to (I2).

- (I2) a. *someone to visit her* \in L2a
 b. i. $\delta(x) \in$ L2a for every $x \in$ L2a, where $\delta(x) =$ *someone to want x*
 ii. $\kappa^2(x, y) \in$ L2a for every $x, y \in$ L2a, where $\kappa^2(x, y) =$ *x and y*
 iii. $\eta(x) \in$ L2 for every $x \in$ L2a, where $\eta(x) =$ *she wants x*
 c. i. if $K \subseteq W^*$ also satisfies (I2a) and (I2bi-ii), then $L2a \subset K$
 ii. if $K \subseteq W^*$ also satisfies (I2biii), then $L2 \subset K$

(G2) $N = \{S, S'\}$

$T = \{\textit{and, her, she, someone, to visit, to want, wants}\}$

$A = \{S\}$

$P = \{S \rightarrow \textit{she wants } S', S' \rightarrow \textit{someone to want } S', S' \rightarrow \textit{someone to visit her, } S' \rightarrow S' \textit{ and } S'\}$

⁵ Although the strings *she wants to visit her* and *she wants someone to visit* are both well-formed in English, neither contains the obligatorily bound empty string in question.

According to (I2) and (G2), every member of L2a with more than two conjuncts is structurally ambiguous. For example the string *p and q and r*, where $p, q, r \in L2a$, is analyzed either as the compound of *p and q* with *r*, or as the compound of *p* with *q and r*. However it fails to analyze that string in the most natural way, as the compound of *p, q, r* directly. The inductive definition (I3) provides that analysis in addition to the other two, as does the production system (G3) obtained from (G2) with the addition of the production $S' \rightarrow S' \text{ and } S' \text{ and } S'$.

- (I3)
- a. *someone to visit her* $\in L3a$
 - b.
 - i. $\delta(x) \in L3a$ for every $x \in L3a$, where $\delta(x) = \textit{someone to want } x$
 - ii. $\kappa^2(x, y) \in L3a$ for every $x, y \in L3a$, where $\kappa^2(x, y) = x \textit{ and } y$
 - iii. $\kappa^3(x, y, z) \in L3a$ for every $x, y, z \in L3a$, where $\kappa^3(x, y, z) = x \textit{ and } y \textit{ and } z$
 - iv. $\eta(x) \in L3$ for every $x \in L3a$, where $\eta(x) = \textit{she wants } x$
 - c.
 - i. if $K \subseteq W^*$ also satisfies (I3a) and (I3bi-iii), then $L3a \subset K$
 - ii. if $K \subseteq W^*$ also satisfies (I3biv), then $L3 \subset K$

However, (I3) and (G3) fail to analyze any member of L3 with more than three conjuncts as the direct compound of those conjuncts, and failure of this sort cannot be eliminated by any inductive definition with finitely many inductive clauses or by any production system with finitely many productions.

A solution for the case of production systems is to replace the set of productions in (G3) that introduce *and* with the *production schema* $S' \rightarrow S' (\textit{and } S')^+$ (Langendoen 1976); call the resulting system G^+ . A solution for the case of inductive definitions is as follows. Let κ^+ be a function that maps any sequence $\sigma = x_1, x_2, \dots$ of two or more members of L^+ to a member of L^+ in the manner of κ^2 and κ^3 . Then replace (I3) by the inductive definition I^+ .

- (I⁺)
- a. *someone to visit her* $\in L^+a$
 - b.
 - i. $\delta(x) \in L^+a$ for every $x \in L^+a$, where $\delta(x) = \textit{someone to want } x$
 - ii. $\kappa^+(\sigma) \in L^+a$ for every sequence σ of two or more members x_1, x_2, \dots of L^+a , where $\kappa^+(\sigma) = x_1 \textit{ and } x_2 \dots$
 - iii. $\eta(x) \in L^+$ for every $x \in L^+a$, where $\eta(x) = \textit{she wants } x$
 - c.
 - i. if $K \subseteq W^*$ also satisfies (I⁺a) and (I⁺bi-ii), then $L^+a \subset K$
 - ii. if $K \subseteq W^*$ also satisfies (I⁺biii), then $L^+ \subset K$

By requiring the set from which the members of L^+ are drawn to be denumerably infinite (e.g., W^*), I^+ defines L^+ to be denumerably infinite, since every sequence over which κ^+ operates is constrained to be finite.

However, nothing in the definition of the range of κ^+ requires it to be a finite sequence.⁶ Suppose that the universal set from which the members of L^+ are drawn is the nondenumerably infinite set W^{**} of strings over W of finite or denumerably infinite length.⁷ From the fact that κ^+ maps every sequence, finite or infinite, of members of L^+ to a member of L^+ , it follows that L^+ is

⁶ Tarski (1935) and Mostowski (1957) pioneered the use of predicates as functions on infinite sequences; see Koslow (1992: 181-182) for discussion. Here we are dealing with a simpler case, since conjunction is not a predicate.

⁷ The cardinality of W^{**} is that of the set of real numbers, i.e., \aleph_1 .

nondenumerably infinite.⁸ Assuming that English is a subset of W^{**} , satisfies (I^+biii) , and contains some member of W^{**} that does not satisfy (I^+biii) , then by (I^+c) , $L^+ \subset \text{English}$, and consequently English is also nondenumerably infinite.

Langendoen and Postal (1984) contend that a production system cannot in principle account for nondenumerable sets, because the latter is limited to generating at most a denumerably infinite set. However, if a production system is understood as a notation for an inductive definition, then (I^+bii) can be expressed in such a system, by replacing the production schema $S' \rightarrow S'$ (and S'^+) by one in which the right side expresses a possibly infinite sequence of two or more S' s with intervening *ands*.⁹

Finally observe that only the inductive clause (I^+bii) has the potential to select members for L^+ from W^{**} that are not also in W^* , i.e., infinitely long members. The inductive clauses $(I1bi)$, $(I2bii)$ and $(I3biii)$ do not. That is, even if we were to enlarge the universal set from W^* to W^{**} in the inductive definitions $(I1)$, $(I2)$ and $(I3)$, the resulting languages $L1$, $L2$ and $L3$ would remain denumerably infinite. Figure 1 shows the relation between the maximum size of a language and the maximum number of arguments in the functions of the inductive clauses of its inductive definition. The actual size may be smaller, depending on the size of the universal set from which the language is drawn.

Maximum number of arguments	Maximum size of language
None (i.e., no induction)	Finite
Finitely many	Denumerably infinite (\aleph_0)
Denumerably infinitely many	Nondenumerably infinite (\aleph_1)

Figure 1. Maximum size of a language in relation to maximum number of arguments in the functions of the inductive clauses of its inductive definition

So far, the only definitive example of an inductive function with at most denumerably infinitely many places for natural languages is the one formulated for (I^+) and its variants. Uriagereka (2005) considers, but does not formalize, another possibility, that the attachment of disjuncts, the class of adjuncts that do not scope over one another, can give rise to infinite sentences, and perhaps more interestingly, to infinitely large forms of interpretation expressible with finite phonologies.

Questioning nondenumerably infinite natural languages

(I^+) is a reformulation of a restricted version of the Closure Principle of Coordinate Compounding of Langendoen and Postal (1984).¹⁰ This reformulation, together with the discussion leading up to it, makes clear exactly what assumptions are needed in order to conclude that English is nondenumerably infinite, namely those in (A') .

⁸ The proof is straightforward. First, the cardinality of L^+ is at least \aleph_1 since L^+ contains a subset which is the size of the power set of the set $L1$, whose cardinality is \aleph_0 . Second, the cardinality of L^+ is at most \aleph_1 since L^+ is a subset of W^{**} . Therefore the cardinality of L^+ is exactly \aleph_1 .

⁹ Pullum and Scholz (2001: 23) cite the work of Thomas (1990) on generating infinite strings in support of their position that Langendoen and Postal's (1984) claim that GES systems are falsified by their failure to describe such strings is "completely misguided". In fact, as Zeitman (1993) more tactfully points out, the corresponding automata-theoretic notions of computation on infinite strings were worked out in the 1960s, and were applied to formal language theory in the 1970s. Langendoen and Postal (1984) cite some of this literature to make the point that the imposition of the finite-length restriction (i.e., taking the universal set to be W^*) is not necessary.

¹⁰ The general form of the Closure Principle of Coordinate Compounding results in the claim that natural languages include members of every transfinite length, not just \aleph_0 and \aleph_1 . I return to this claim below.

- (A') a. English satisfies (I^+a) and (I^+b) .
 b. The correct choice of universal set of which English is a subset is non-denumerably infinite (e.g., W^{**}).

Concerning (A'a), it can be reasonably objected that the inductive clause (I^+b) as it stands is not correct. Some minor objections are dealt with in Langendoen and Postal (1984). The more recent efforts to analyze coordinate compounding as a kind of right-branching subordination are no more successful at characterizing the fundamental structural properties of coordination than (G2) above, see Langendoen (1998) for discussion. A more serious objection is based on the finding in Langendoen (1998) that the degree of embedding of coordinate compounds within coordinate compounds is finitely bounded. To overcome this objection, the definition of L^+ is introduced into the definition of L^+ as an intermediate step, and (I^+bii) , is restricted in its operation to members of L_1 , as in (I^+') .

- (I^+') a. *someone to visit her* $\in L^+a$
 b. i. $\delta(x) \in L^+a$ for every $x \in L^+a$, where $\delta(x) = \textit{someone to want } x$
 ii. $\kappa^+(\sigma) \in L^+b$ for every sequence σ of two or more members x_1, x_2, \dots of L^+a , where $\kappa^+(\sigma) = x_1 \textit{ and } x_2 \dots$
 iii. $\eta(x) \in L^+$ for every $x \in L^+b$, where $\eta(x) = \textit{she wants } x$
 c. i. if $K \subseteq W^{**}$ also satisfies (I^+a) and (I^+bi) , then $L^+a \subset K$
 ii. if $K \subseteq W^{**}$ also satisfies (I^+bii) , then $L^+b \subset K$
 iii. if $K \subseteq W^{**}$ also satisfies (I^+biii) , then $L^+ \subset K$

With this change, all embedding of coordinate compounds is excluded from members of L^+ without affecting L^+ 's status as nondenumerably infinite.

Turning now to (A'b), the assumption has not so much been argued against, as simply “dismissed”, as Hinzen and Uriagereka (2006: 91) recently stated, and the view that natural languages contain more than denumerably many sentences has been described as “esoteric” (Dale 1996: 99). Although infinite strings and structures, and computations over them have been widely investigated in logic and computer science, the need for them in linguistics has not been seen to be compelling. However, as pointed out by Pullum and Scholz (2005: 17), Schiffer (1972) and afterwards Joshi (1982) formulated the notion of mutual belief as the infinite conjunction of propositions such as *Jones believes that iron rusts, and Smith believes that iron rusts, and Jones believes that Smith believes that iron rusts, and Smith believes that Jones believes that iron rusts, and Jones believes that Smith believes that Jones believes that iron rusts, and ...*, and other uses for infinitely long natural-language expressions may eventually be found.

Justifying nondenumerably infinite natural languages

Pullum and Scholz (2005: 18) observe that it is not clear why a sentence like the infinite expression of mutual belief given above should be considered ungrammatical, “since the string is entirely Englishlike in terms of its grammatical properties. But if it is, then some English expressions have infinite length.” This echoes the point made in Langendoen and Postal (1984: 42ff.) that Katz’s argument cited above in the discussion of the justification of denumerably infinite natural languages applies equally well to expressions with infinite length. It also comports with Zeitman’s conclusion that the “most compelling argument for the study of infinite sentences is that the linguistically relevant patterns of finite ones occur also in the infinite case.... The actual examples Langendoen and Postal give of this sameness of patterns is precisely the

kind of sameness of patterns that occurs when one extends the finite string automata and grammars to their infinite string versions.” (1993: 35)

Can natural languages have more than nondenumerably many members?

Langendoen and Postal (1984) interpret their Closure Principle of Coordinate Compounding as implying that there are natural language expressions of every transfinite length, and that the resulting collection is too big to be considered a set. Pullum and Scholz (2005: 18) question this interpretation on the grounds that it is “not just unmotivated but actually unstatable” in Model-Theoretic Syntax theory. Nor can such a collection be characterized inductively or by a production system. To see this, note that if the universal set over which I^+ operates is specified as W^{***} , the power set of W^{**} , whose cardinality is \aleph_2 , the inductively defined language does not expand to take advantage of this new capacity; it remains \aleph_1 . A new inductive definition with a new inductive clause with a function that ranges over as many places as there are real numbers would be required to do so. Even if we could find a linguistic motivation for such a definition, which we do not have at the moment and presumably never will, it would not be capable of making use of the space of possibilities opened up by selecting W^{****} , the power set of W^{***} , as the universal set. And so forth. I conclude that the chart in Figure 1 exhausts the range of possible sizes for natural languages.

References

- Dale, Russell Eliot (1996). The theory of meaning. Ph.D. Dissertation, City University of New York. <https://webspace.utexas.edu/deverj/personal/test/theoryofmeaning.pdf> (last visited 2007-04-01).
- Everett, Daniel (2007). Cultural constraints on recursion. This volume.
- Hintzen, Wolfram and Juan Uriagereka (2006). On the metaphysics of linguistics. *Erkenntnis* 65:71-96.
- Hopcroft, J. E. and J. D. Ullman (1979). *Introduction to Automata Theory, Languages and Computation*. Reading, MA: Addison-Wesley.
- Joshi, Aravind (1982). Mutual beliefs in question-answer systems. In *Mutual Knowledge*, Neil Smith (ed.), 181-197. London: Academic Press.
- Joshi, Aravind (2007). Does recursion in language work the same way as in formal systems? This volume.
- Katz, Jerrold J. (1966). *The Philosophy of Language*. New York: Harper and Row.
- Koslow, Arnold (1992). *A Structuralist Theory of Logic*. Cambridge: Cambridge University Press.
- Langendoen, D. Terence (1976). On the weak generative capacity of infinite grammars. *CUNYForum* 1:13-24.
- Langendoen, D. Terence (1998). Limitations on embedding in coordinate structures. *Journal of Psycholinguistic Research* 27:235-259.
- Langendoen, D. Terence and Paul M. Postal (1984). *The Vastness of Natural Languages*. Oxford: Basil Blackwell.
- Lasnik, Howard (2000). *Syntactic Structures Revisited: Contemporary Lectures on Classic Transformational Theory*. Cambridge, MA: MIT Press.
- Mostowski, A. (1957). On a generalization of quantifiers. *Fundamenta Mathematicae* 44:12-36.

- Pullum, Geoffrey K. and Barbara C. Scholz (2001). On the distinction between model-theoretic and generative-enumerative syntactic frameworks. In *Logical Aspects of Computational Linguistics: 4th International Conference*, P. de Groote, G. Morrill and C. Retoré (eds.), 17-43. Berlin: Springer.
- Pullum, Geoffrey K. and Barbara C. Scholz (2005). Contrasting applications of logic in natural language syntactic description. In *Logic, Methodology and Philosophy of Science 2003: Proceedings of the 12th International Congress*, Peter Hajek, Luis Valdes-Villanueva and Dag Westerthal (eds.), 475-496. London: KCL Publications. <http://www.ling.uni-potsdam.de/~rvogel/grundlagen/PS05.pdf> (last visited 2007-04-01; page references are to the on-line version).
- Schiffer, Steven R. (1972). *Meaning*. Oxford: Clarendon Press.
- Tarski, Alfred (1956 [1935]). The concept of truth in formalized languages. In *Logic, Semantics and Metamathematics*, J. H. Woodger (trans.). Oxford: Oxford University Press. Previously published as Der Wahrheitsbegriff in den Formaliserten Sprachen. *Studia Philosophica* 1:261-405.
- Thomas, Wolfgang (1990). Automata on infinite objects. In *Handbook of Theoretical Computer Science*, J. van Leeuwen (ed.), 135-191. New York: Elsevier.
- Uriagereka, Juan (2005). Adjunct space? Unpublished paper presented at the Prospects for Dualism conference, Amsterdam.
- Zalabardo, José A. (2000). *Introduction to the Theory of Logic*. Boulder, CO: Westview Press.
- Zeitman, Suzanne (1993). Somewhat finite approaches to infinite sentences. *Annals of Mathematics and Artificial Intelligence* 8(1-2):27-36.