J Log Lang Inf
DOI 10.1007/s10849-014-9203-2

Resolving the Infinitude Controversy

András Kornai

Abstract A simple inductive argument shows natural languages to have infinitly many sentences, but workers in the field have uncovered clear evidence of a diverse group of 'exceptional' languages from Proto-Uralic to Dyirbal and most recently, Pirahã, that appear to lack recursive devices entirely. We argue that in an information-theoretic setting non-recursive natural languages appear neither exceptional nor functionally inferior to the recursive majority.

Keywords Recursion · Information content

1 Introduction

The central topic of the infinitude debate is a putative syntactic universal:

$$\textit{natural languages have infinitely many grammatical sentences.} \qquad (1)$$

Though the debate is often framed in terms of recursion rather than infinitude, we will deliberately avoid talking about recursion in this context, given the confusing 'legacy of imprecision' left by Chomsky (1957) in this domain, and develop an information-theoretic perspective where the more precise distinctions rightfully urged by Tomalin (2011) will play only a secondary role. This is not to say that Tomalin succeeded in banishing imprecision from this discussion once and forever. For example Watumull et al. (2014), while seemingly accepting the stan-

For Marcus Kracht on his 50th birthday.

A. Kornai (✉)
HAS Computer and Automation Research Institute, Kende u 13-17, Budapest 1111, Hungary
e-mail: kornai@math.bme.hu; andras@kornai.com

Published online: 15 July 2014

dard definition of recursiveness, take what is an upper bound for natural language stringset complexity to be a necessary condition, arriving at their conclusion by *petitio principii*:

> A computable function is necessary to derive sets; a list is epistemologically and, more importantly, ontologically meaningless. [...] Support for rejection of the [recursivity] thesis speaks directly to our conception of human nature (i.e., those traits unique to and definitional of the species) and the nature of the universe (i.e., the ontological interconnection of mathematics and biology).

In fact, given Tomalin's argument that "(...) in Chomsky's theoretical framework, anything that permits the generation of an infinite set of grammatical sentences can be referred to as being 'recursive'," whatever we say here about 'infinitude' here will apply, without change, to the 'recursion' debate, and we will speak of recursive, iterative, and 'looping' devices indiscriminately as these can all generate infinite stringsets.

What gives the infinitude debate particular importance is that in general the question whether a given structure has a finite or an infinite number of elements greatly impacts the range of formal techniques that can be brought to bear. On the one hand, exhaustive listing works very well for finite sets, while infinite ones must be characterized by more complex means. On the other, analytic techniques based on limit processes are very powerful for the infinite case, but yield little that is useful when applied to finite sets. Therefore, an early decision on whether a given structure with a large number of elements is better approached from the finite or the infinite side would have great heuristic value, and it should come as no surprise that the issue has received considerable attention in every branch of linguistics.

Before the emergence of the phonemic principle, many linguists thought they have to confront an infinite variety of speech sounds continuously blending into one another, and it took a great deal of conceptual work to tease phonetics and phonology apart (Anderson 1985). In morphology, there is still not a firm consensus whether there are finitely many words. In syntax, it is clearly the infinite view that has more adherents, but time and again we come across languages where only a finite presentation seems to make sense: the list includes Dyirbal (Dixon 1972), Walbiri, Wargamay, and other Australian languages, Hixkaryána (Derbyshire 1979) and other Amazonian languages, Proto-Uralic (Ravila 1960), archaic Chinese, early Akkadian, and other ancient languages.

Recently, the infinitude debate (for a critical overview, see Pullum and Scholz 2010) has been rekindled by the publication of Everett (2012). For example Bartlett (2012) writes: "[Pirahã] doesn't follow one of the fundamental tenets of linguistics, a finding that would seem to turn the field on its head, undermine basic assumptions about how children learn to communicate". In this paper we steer clear of the narrower debate surrounding the empirical facts and their competing interpretations (Nevins et al. 2009; Everett 2009; Sauerland 2010; Piantadosi et al. 2012)—our focus is with the theoretical impact, if any, of Pirahã exceptionality. We will argue that from an information-theoretical perspective there is no bright line between finite and infinite languages, and there isn't even a fuzzy line: the information-carrying capacity of a finite language F can exceed that of an infinite language R.

In Sect. 2 we briefly present the standard arguments in favor of the infinite view. In Sect. 3 we discuss the dependence between recursion and a simple counting measure, *average sentence length*. In Sect. 4 we present our main argument using the strongly related notions of frequency and information content. By applying the classic (Shannon 1948) measure of informativeness, the infinitude debate shifts from a naive count measure of *how many distinct messages* to a more nuanced measure of *how much information* can be supported by a finite or infinite language. By recasting the infinitude issue in information-theoretic terms the problem disappears: Pirahã, Dyirbal, Hixkaryána, and similar languages with finite syntax, are no longer exceptional.

2 The Standard View

The standard view, familiar both from textbooks and research papers, is generally stated as (1) or, equivalently, as a claim of No Maximal Length (NML)

For any English expression there is another expression that is longer. (2)

This position, originating with Chomsky (1957), is based on two far-reaching methodological observations. First, we know that any corpus we can think of, such as the sentences heard by a child during language acquisition, the sentences committed to writing, or even the totality of the sentences produced by speakers of a given language, is finite. However, these are just samples from a larger population, and it in no way follows from the finiteness of samples that the population itself is finite.

> Any grammar of a language will *project* the finite and somewhat accidental corpus of observed utterances to a set (presumably infinite) of grammatical utterances. (Chomsky 1957:15)

To drive the point home, one may consider the set of natural numbers, which we know to be infinite, yet all numbers used in all computations by all humans and computers until now (and given standard assumptions about the finite lifespan of the universe, ever) fit in a finite set. What really matters here is the set of *potentially* usable numbers, given to us by a generative system that includes a successor operation generating new numbers from old, with an axiom to guarantee that the process never ends. This much is hardly controversial, and serves quite well to defend the infinitude hypothesis against the most naive attacks. More important, it shifts attention from the corpus to where it belongs, the language itself. The second observation is a bit more subtle.

> In general, the assumption that languages are infinite is made in order to simplify the description of these languages. If a grammar does not have recursive devices (...) it will be prohibitively complex. If it does have recursive devices of some sort, it will produce infinitely many sentences. (Chomsky 1957:23–24)

The current round of the infinitude controversy stems from Everett's observation that Pirahã, as far as he can see, simply lacks recursive (and even the weaker kind of itertive/looping) devices. Given the logic of Chomsky's argument, this is not particularly damning—even if Pirahã lacks recursion, iteration, and looping, there are plenty of languages that have these. Yet there is quite a bit of consternation surrounding Everett's findings, for two main reasons. First, the existence of such 'exceptional' languages directly contradicts the extremist position taken in Hauser et al. (2002) that recursion constitutes the *only* uniquely human, language-specific part of the language faculty. Fortunately, more nuanced views of the human language faculty and its evolution are not particularly threatened by this (see Jackendoff and Pinker 2005). Second, the existence of 'exceptional' languages calls into question a kind of argument that has long been standard (see Pullum and Scholz 2010) in generative linguistics. For example, Lakoff (1968:5) writes in her thesis

> (...) the traditional view (assumes) the (Latin) proto-language could not have complex sentences. (...) If this assumption were realistic, and the proto-language actually could not embed sentences inside others, it could easily be shown that this proto-language had only a finite number of sentences, unlike any natural language known to linguistics.

Again, the loss is far from catastrophic, since it is only indirectly, mediated by (1), that the existence of finite languages like Pirahã could impact our analysis of proto-Latin. If (1) fails, we may still think proto-Latin had recursion, but the evidence now will have to come from subordination elsewhere in Romance and in IE, the convenient counting argument is no longer available.

3 Sentence Length

We assume, as is standard, a finite alphabet V and a set $L \subset V^*$. When we think of members of V as phonemes, the finiteness assumption is not particularly limiting, since phonemic inventories are on the order of 10^1–10^2. However, when we think of the letters of the alphabet as fully formed words, the hypothesis $|V| < \infty$ baked into the formalism takes us directly to the heart of Chomsky's argument: to the extent word-formation involves looping processes such as *anti-* or *grand-* prefixation or noun–noun compounding, the number of potential words is infinite (see Langendoen 1981; Kornai 2002). Nominalization and incorporation processes that appear to feed back syntactic structures into the word formation component of the grammar are of particular interest in this regard. To avoid prejudging the issue, we will take the letters in V to be preterminals (lexical categories, part of speech tags) whenever necessary, so we will look at Adj Adj N.PL $\mathrm{V_I}$.3SG Adv instead of *colorless green ideas sleep furiously*—sentence length is not affected by this, but the overall picture is much simplified.

Let us first take L to be English, a language that is not even suspected of being finite. Therefore, as we take ever-increasing samples $S_1, S_2, \ldots$ from L, we expect to see longer and longer sentences. Indeed, the first sample from the British National Corpus (BNC, see http://www.natcorp.ox.ac.uk) comprising 52k sentences (844k words), we already find a very readable sentence of 334 words, or 377, if we count punctuation tokens as separate words. (This is commonly done in computational linguistics, and all data presented in this paper will follow this tokenization convention.) As we shall see shortly, large corpora support quite clearly claim (2) in that they have no natural cutoff-point: longer sentences may be rare, but the grammar itself should be prepared for arbitrary sentence length.

On average, however, English sentences are much shorter. The BNC is divided into 175 subcorpora ranging from 7k to 2m words, and showing average sentence lengths from 7.1 to 32.2. The grand average is 18 words per sentence in the BNC, and 10.8 words/sentence in the much larger Google webcorpus (W1T, see Brants and Franz 2006), containing statistical summaries of English web text totaling 1T (10^{12}) words. The smallest average sentence length is found in transcriptions of spoken material, the largest in legal texts.

When we count the number of words per *subsentence* (comma-separated stretches, without analyzing in detail whether these are clauses, phrases, parentheticals, or even typos), average length drops to 10.5, quite close to the average sentence length in spoken materials. In particular, the legal material that was in the lead before has only 9.6 words per subsentence, while spoken material transcribed by slightly different conventions regarding the placement of commas can go as high as 13.7 words per subsentence. The 300+ word BNC sentence mentioned above has an average subsentence length of 8.8, which is well within the one sigma range for subsentences, and the comparable W1T datum of 7.2 is only slightly below one standard deviation.

The BNC is two orders of magnitude larger than the classic Brown corpus, but it does not substantially change, let alone invalidate, the classic results obtained on the smaller corpus by Francis and Kučera (1982:552). The results presented here are not strictly comparable to theirs, in that Francis and Kučera use a sophisticated regular expression to extract subsentence-level units they call *predications*, while we simply counted punctuation, but the basic statistics are unchanged: they see 6.6 words per predication in fiction, 8.6 in nonfiction (the numbers are from their Table 6.2, incremented by 1 so as to account for the punctuation tokens).

As we shall see, sentence length distribution is far from normal (the heavy tail is quite visible in density plots), so using standard deviation to measure departure from the norm could be misleading. Since the departure is toward heavier than normal tail, this actually reinforces the main point these numbers drive home, that an average measure can stay within small bounds, even when the population contains *more* than the normally expected number of arbitrarily large examples.

On a broader sample of web corpora from 18 language varieties (see Zséder et al. 2012) ranging from a low sample size N_s of 1.45 million sentences (Nynorsk) to a high of 110.79 million (Dutch) we find very similar results, see Table 1. Average sentence length L goes from a low of 14.8 (Finnish) up to 34.5 (Serbian), and in general the variance σ_L exceeds the mean length L, indicating a heavy tail. Average subsentence length U is far more constrained, from a low of 7.3 (Finnish) up to 10.3 (Spanish), and again the variance σ_U exceeds the mean U. The number of subsentences per sentence U/L goes from a low of 1.76 (Dutch) up to 3.23 (Serbian).

It is worth noting that many of the extremes are found in the small Nynorsk and Serbian samples—the expectation is clearly that the key statistics of larger samples will regress toward the mean, which appears to be about 20–21 words per sentence (with a cross-language variance of 4–5), 8.8 words per subsentence (cross-language

Table 1 Sentence and subsentence length in various languges

Language	N_s	L	σ_L	N_U	U	σ_U	U/L	$\sigma_{U/L}$
Catalan	25.80	25.7	34.9	62.88	10.0	13.6	2.44	2.67
Croatian	66.85	22.5	33.0	158.21	8.9	12.2	2.37	3.04
Czech	33.15	18.7	25.1	77.88	7.4	9.2	2.36	2.89
Danish	28.81	17.4	26.3	56.47	8.4	12.8	1.97	2.29
Dutch	110.79	18.1	25.6	193.80	9.9	13.0	1.76	1.99
Finnish	61.54	14.8	20.1	118.16	7.3	8.8	1.93	2.22
Indonesian	13.96	22.4	31.3	31.87	9.2	12.3	2.29	3.08
Lithuanian	86.62	16.4	26.3	199.60	6.6	9.8	2.31	2.58
Norwegian (nn)	1.45	18.2	17.3	2.60	9.7	8.8	1.80	1.64
Norwegian (no)	90.19	18.2	24.7	167.34	9.3	12.0	1.86	2.10
Polish	77.48	18.6	33.8	174.53	7.8	13.8	2.27	3.37
Portuguese	40.89	23.8	40.1	100.82	9.0	15.0	2.47	3.25
Romanian	38.71	27.6	44.1	104.64	9.6	16.1	2.70	3.82
Serbian (sh)	38.24	22.3	29.7	92.42	8.6	11.6	2.42	2.72
Serbian (sr)	2.23	34.5	70.7	7.19	10.0	16.1	3.23	5.28
Slovak	42.70	20.4	36.1	100.41	8.1	14.0	2.36	3.15
Spanish	51.18	27.5	44.5	128.80	10.3	18.2	2.53	3.27
Swedish	56.67	15.9	24.5	94.44	9.1	12.7	1.69	1.94

variance slightly above 1), and 2.25 subsentences per sentence (variance 0.4). Francis and Kuˇcera have 2.64 predications per sentence in the Brown Corpus. Both their results and the BNC data, with an average sentence length of 18, subsentence length of 10.5, and subsentence per sentence ratio 1.83, fits quite well with the crosslinguistic picture presented above. We have no access to machine readable data, but from what has been reported in the literature, Pirahã seems to differ from English not so much in average subsentence length as in the average number of subsentences per sentence, a matter we shall return to in Sect. 5.

While 9–10 words (the 10.5 average includes punctuation) may seem too short, it is very hard to find any problematic constructions of English that we cannot illustrate in an example with fewer than ten words. Wh-extraction out of negated comparative complement? *Name someone you're not cleverer than*. Across-the-board topicalization? *This, nobody saw and nobody heard*. Stacked relative clauses? *Find someone you despise that I hired*. It seems that if we could describe all English sentences of up to nine words in length, we would have no remaining descriptive problems at all.

The empirical basis of our claim of small average sentence and subsentence length, resting on several billion words of non-English text and over a trillion words of English, is rather strong: as we go to larger samples, average sentence length does not grow. To gain a better theoretical understanding, consider the example of a simple probabilistic grammar containing only two states, corresponding to the outcome of tossing a fair coin. In one state, the grammar outputs the string *grandfather was a true pioneer* and in the other it outputs *great-*. A moment of thought will show that the grammar generates with probability 1/4 *great-grandfather was a true pioneer*, with probability 1/8 *great-great-grandfather was a true pioneer* and in general with probability $1/2^{n+1}$ *great^n grandfather was a true pioneer*. It is thus a truly looping grammar, generating arbitrarily long sentences, but no matter how large a random sample we take, average sentence length will stay around 6. Further, if we decide that we are quite happy with a grammar that characterizes 99.9 % of the data, we can exclude from consideration every sentence with 9 or more repetition of *great*s, and thus forcibly render the system finite. Yet this actually complicates the system, since now we need to set up some kind of regulator mechanism that counts up to nine and stops the generation if that limit is reached. The number of states is a standard measure of automaton complexity: by this measure the original system had only 2 states (corresponding to heads and tails), the modified one has at least 10.

In the process of limiting the system to short output it seems it is not just the simplicity of the grammar that has been lost, but also something even more important, the capacity of using the system to convey an infinite variety of meanings corresponding to the infinite variety of situations we may wish to describe as communicating agents. Clearly, *grandfather* and *great-grandfather* does not mean the same thing, the former is the son of the latter, and the same relation holds between *great-grandfather* and *great-great-grandfather*, *great-great-grandfather* and *great-great-great-grandfather*, so all the above sentences mean something different. With one grammar, we can express an infinitude of meanings, and with the other, curiously even more complicated grammar, we can only express a finite variety. If the loss is truly this momentous, if indeed we are crossing a bright line between human and animal communication, we begin to understand why the infinitude debate is so heated.

4 Information Content

To see how much difference the finite/infinite distinction actually makes to communicative ability, we need to recall the basic model of communication originally introduced by Shannon (1948): there is a *sender*, there is a *receiver*, and there are *messages*, finitely or infinitely many, that can pass between them. (Shannon actually considered the case of messages getting corrupted en route, hence the name *noisy channel model*, but we will stay in a noise-free setting here.) Readers will know that the information transmitted through the channel is measured in *bits*, and may recall the formula $H = -\sum_i p_i \log p_i$ determining the maximum capacity of the channel. It is important to keep in mind that H is a pure measure of capacity: when we say that a modem line can carry 10 megabits per second, this says nothing about the nature of the information in these bits, just as when we say that a truck can carry 10 tons this says nothing about the nature of the items that make up the load.

What is the information carrying capacity of our example language? It doesn't matter that the messages describe my grandfather, my great-grandfather, my great-great-grandfather and so on. In fact these may even be coded secret messages to my broker actually meaning 'don't buy Exxon stock', 'buy one share of Exxon', 'buy two shares', and so on. But as long as these messages are sent once a second and follow the probability distribution $1/2, 1/4, 1/8, \ldots$ their information content is just 2 bits per second. There may be an infinite variety of these messages, but the information carried by them *on the average* is still a finite number, for the exact same reason why sentence length averages can be finite in spite of the appearance of arbitrarily long sentences. There can be tremendously informative messages, as when I ask my broker to buy exactly 37,272 shares of Exxon, but they appear very rarely, and are swamped out by the less informative shorter messages. When we artificially truncate the distribution as in the example above, we indeed lose some information carrying capacity, going to 1.976 bits from 2.

The same phenomenon can be seen in the distribution of words. English words carry on the average some 12.7 bits of information, Hungarian words, being composed of more morphemes, carry 15.4 (see Kornai 2008:Ch.7 for how these numbers can be computed from frequency counts). We don't yet have large Pirahã corpora, but we note here that one can surpass the informativeness of the English vocabulary by a closed list of only 7,000 words, provided these seven thousand were equiprobable. While the difficulties in getting reliable estimates multiply as we go from words to sentences, the principles are unchanged, and *it is simply not true that a finite language must have smaller information carrying capacity than an infinite one*. For example, the language $\{1, 2, 3, 4, 5, 6\}$ that we use to report the outcome of tossing a fair die has only 6 utterances, but with $H = 2.585$ it is almost 30 % more informative than the infinite language of our example with $H = 2$. Altogether, as a communications device a language that lacks any form of recursion (iteration, looping) need not be in any way inferior to one that has one or more of these. In place of (1), we must propose a far more modest universal:

natural languages are communications devices of finite capacity. (3)

Chomsky (1965) would have called (3) a substantive, as opposed to a formal, universal: it is not at all the case that any formal language whose strings s_i appear with some prescribed[1] probabilities p_i will have a finite entropy H. For a counterexample, consider a language N where the probability of the kth string (in lexicographic ordering) is $1/\log_2(k+1) - 1/\log_2(k+2)$—it is trivial to see that the p_k sum to 1, and it requires only a simple argument (see Baer 2013 ms) to demonstrate that entropy is divergent.

It is a bit harder to speculate on what Shannon would have said about (3), especially as his framework predates Chomsky's early work on formal languages by nearly a decade. Clearly, he considered natural language to be very much in scope for the noisy channel model, and likely he would have considered N a degenerate case, in that his primary interest was with telephone, telegraph, and other devices of finite capacity. The mere fact that he experimented with the estimation of character and

[1] According to some, "in human language communications, the probability of an utterance varies from situation to situation, moment to moment: if an elephant appears on the university campus, this affects the probability of 'elephant'-utterances, threatening the empirical basis of (3)" This view rests on a confusion between the probability value and the method of sampling: clearly average human height is not at all affected by the fact whether we use the basketball team or the kindergarten as our sample, it's just that neither sample is very representative.

word entropy can be seen as offering some indirect support for the view that he would have endorsed (3).

In this regard, it should be emphasized that the information carrying capacity of human languages is not just finite but quite puny compared to modern telecommunications networks. One second of human conversation carries *at most* a few hundred bits at the acoustic level (Padellini and Capman 2004), and at most a few dozen bits at the symbolic level (Brown et al. 1992), while a fiber optic cable can carry billions of bits. Were this not the case, we could not cram thousands of simultaneous conversations into a single transatlantic cable, just as we cannot cram thousands of elephants into a single ten ton truck.

5 Conclusions

People on both sides of the infinitude controversy agree that the finite case, exemplified by Pirahã (and quite possibly by Dyirbal, Walbiri, Wargamay, Hixkaryána, and several reconstructed languages) is far more rare than the infinite case that we see in English, where a statement like (2) can be substantiated by many constructions involving direct and indirect quotation, attitudinal verbs, etc. not to speak of coordination, relativization, and other constructions that will all give rise to infinite stringsets. Given that animal communication systems appear to lack recursion (there is some evidence of more mechanistic repetition of e.g. mating calls one may consider iterative/looping behavior), it is rather tempting to say (Hauser et al. 2002) that distinctively human communication must evolve from 'primitive' (non-recursive, non-looping, non-iterative) finite systems to somehow more 'advanced' (recursive/looping/iterative) infinite ones.

A syntactic mechanism such as coordination serves to make communication less redundant: instead of *I caught a fish. I caught a bird*, we can say *I caught a fish and a bird*, using a single sentence without repeating the subject and the main verb, enabling faster communication. But a language like Pirahã that lacks coordination can of course communicate the same meaning, and what was won by omitting a few words may have been lost in terms of intelligibility.

What needs to be kept in mind in doing these comparisons is that the stakes are not high: fast talkers routinely say 300 words per minute (the world record is over 600), while the comfortable average is around 150. Since the change in speech rate can routinely accommodate doubling (and in exceptional cases, quadrupling) of informativeness, the fraction of a percent gain brought by eliminating these and similar redundancies (e.g. in gapping constructions) will not give any noticeable communicative advantage to languages that permit such constructions over those that do not. Further, the function of reduction, such as omitting optional funtion words like *that* before relative clauses, may not even be the out and out decrease of redundancy, but rather keeping information density relatively even (Levy and Jaeger 2007; Jaeger 2010).

Whether languages evolve toward higher channel capacity is unclear: abbreviatory devices are common to languages, but so are ones that enhance redundancy. But even if there was a clear tendency toward increased informativeness, a rather dubious assumption given the biological limits to human information-processing capacity, this would still fail to create evolutionary pressure toward grammars that *must* generate infinite languages, since finite languages can have more information carrying capacity than infinite ones.

In Sect. 2 we quoted Chomsky's argument to the effect that assuming recursion can simplify the description of languages, and in Sect. 3 we constructed a toy example that shows the phenomenon quite clearly. But the remainder of the argument, that grammars without recursion will be prohibitively complex, assumes what needs to be proven, the presence of recursive (looping, iterative) constructions in the language. If there are cases where limiting the depth of recursion (number of loops, number of iterations) creates unnecessary complexities, there are also cases of the opposite, where *not* limiting recursion (looping, iteration) would create unnecessary complexity. Both are easily exemplified on English prenominal modifiers. A recursive rule like N → A N yields strings like *dog, white dog, big white dog,* etc., permitting the stacking of any number of adjectives modifying the noun. But a similar rule N → Q N would permit the stacking of quantifiers, yielding nominal expressions like **every some dog* which would now have to be excluded by some special mechanism.

Altogether, whether a recursive (looping, iterative) device is useful must be decided on a case by case basis, and rules having no or limited recursion are quite often warranted. The vast majority of the rules in any grammar are lexical (often restricted to individual items) and thus lack any iterative aspect. Let us say the remaining rules each are looping (iterative, recursive) with some probability p: in a system of n rules we are likely to find np such rules. As long as n and p are not too low, most languages will turn out to be infinite because even one such rule is sufficient to make the stringset expand without limits, but as our sample of known languages grows, sooner or later we will find an example of an entirely non-recursive grammar, just as in playing bridge one will sooner or later encounter a hand composed entirely of minor suits. Pirahã is simply the latest, and in no way exceptional, example of the phenomenon.

Counting individual rules and ignoring their interactions simplifies matters somewhat, since individually non-recursive (non-iterative, non-looping) rules such as transitions in a finite automaton can combine to provide recursive (iterative, looping) rule sequences. If this phenomenon is taken into account, the probability of obtaining a grammar that will generate an infinite stringset will be even higher, a matter we can can numerically estimate by taking finite automata to be random graphs in the sense of Erdős and Rényi (1960). We call an automaton *trimmed* if it contains no state that lacks a directed path from the start state or a directed path to an accepting state. Obviously, the yield of the automaton is unchanged by removing (trimming) all such states. Consider automata that have, after trimming, n states and M transitions, possibly including self-loops. The smallest one with nonempty yield will be a chain with $M = n - 1$, and the largest one with nonempty but finite yield will have $M = n(n - 1)/2$ edges (running from state i to state j iff $i < j$, with the initial state numbered 1 and the final n). But on the average it takes very few edges after the first n to guarantee a loop (Łuczak and Seierstad 2009), and an elementary argument shows that in the range of interest, with one language in a thousand being 'exceptional', it takes only ten extra transitions to guarantee that 99.9 % of the grammars will have an infinite yield.

Let us now summarize the argument. The fact that languages like Pirahã lack looping/iterative/recursive constructions like coordination does in no way make them communicatively inferior, since the same meaning remains expressible, just not within the bounds of a single sentence. As we discussed in Sect. 2, the average number U/L of subsentences (predications) per sentence is above 2 in written English, but only slightly above 1 in spoken English. Unsurprisingly, Pirahã resembles spoken English more than it resembles written English. Since average subsentence length is around 10 in English, it is truly tempting to consider a grammar of the English main clause, characterizing all grammatical structures of length $1, 2, \ldots, 10$, (and maybe 11, 12, and 13, just to be on the safe side) and compare this to the structure of Pirahã. Sadly, we cannot make the comparison, because we don't have the requisite English grammar. It is not Pirahã that is causing a seemingly unsolvable problem to the current flavor of generative syntax, it is the vast body of evident, massively documented, and clearly replicable findings about English, Chinese, and all the world's major languages that it cannot deal with. If we cannot account for the short sentences, worries about arbitrarily long ones are at best premature, at worst delusional.

Acknowledgments We thank Geoff Pullum, Christina Behme, and the anonymous referees for detailed comments on earlier versions. Work supported by OTKA Grants #82333 and 77476 and by the European Union and the European Social Fund through project FuturICT.hu (Grant Number TAMOP-4.2.2.C-11/1/KONV-2012-0013).

References

Anderson, S. R. (1985). *Phonology in the twentieth century: Theories of rules and theories of representations*. Chicago: University of Chicago Press.

Baer, M. (2013). A simple countable infinite-entropy distribution. ms. Accessed July, 2013 from https://hkn.eecs.berkeley.edu/~calbear/research/Hinf.pdf.

Bartlett, T. (2012, March 20). Angry words. Chronicle of Higher Education.

Brants, T., & Franz, A. (2006). *Web 1T 5-gram Version 1*. Philadelphia: Linguistic Data Consortium.

Brown, P., Pietra, S. D., Pietra, V. D., Lai, J., & Mercer, R. (1992). An estimate of an upper bound for the entropy of English. *Computational Linguistics*, *18*(1), 31–40.

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Derbyshire, D. (1979). *Hixkaryana*. Lingua descriptive series 1.

Dixon, R. M. (1972). *The Dyirbal language of North Queensland*. Cambridge: Cambridge University Press.

Erdös, P., & Rényi, A. (1960). On the evolution of random graphs. *Magyar Tudományos Akadémia Matematikai Kutató.*, *5*, 17–61.

Everett, D. (2009). Pirahã culture and grammar: A response to some criticism. *Language*, *85*(2), 405–442.

Francis, W. N., & Kučera, H. (1982). *Frequency analysis of English usage*. Boston: Houghton Mifflin.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how does it evolve? *Science*, *298*, 1569–1579.

Jackendoff, R., & Pinker, S. (2005). The nature of the language faculty and its implications for evolution of language (reply to Fitch, Hauser, and Chomsky). *Cognition*, *97*, 211–225.

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, *61*, 23–62.

Kornai, A. (2002). How many words are there? *Glottometrics*, *2*(4), 61–86.

Kornai, A. (2008). *Mathematical linguistics*. Berlin: Springer.

Lakoff, R. (1968). *Abstract syntax and Latin complementation*. Cambridge, MA: MIT Press.

Langendoen, T. (1981). The generative capacity of word-formation components. *Linguistic Inquiry*, *12*(2), 320–322.

Levy, R., & Jaeger, T. (2007). Speakers optimize information density through syntactic reduction. *Advances in Neural Information Processing Systems*, *19*, 849–856.

Łuczak, T., & Seierstad, T. G. (2009). The critical behavior of random digraphs. *Random Structures and Algorithms*, *35*, 271–293.

Nevins, A., Pesetsky, D., & Rodrigues, C. (2009). Piraha exceptionality: A reassessment. *Language*, *85*(2), 355–404.

Padellini, M., Capman, F., & Baudoin, G. (2004). Very low bit rate (VLBR) speech coding around 500 bits/sec. In *Proceedings of the XII. European signal processing conference*, pp. 1669–1672.

Piantadosi, S.T., Stearns, L., Everett, D.L., & Gibson, E. A. (2012). Corpus analysis of Pirahã grammar: An investigation of recursion. LSA annual meeting. Accessed December, 2013 http://tedlab.mit.edu/tedlab_website/researchpapers/Piantadosi_et_al_2012_LSAtalk_Piraha.pdf.

Pullum, G. K., & Scholz, B. C. (2010). Recursion and the infinitude claim. In H. van der Hulst (Ed.), *Recursion in human language (studies in generative grammar)* (pp. 113–138). Berlin: Mouton de Gruyter.

Ravila, P. (1960). Proto-Uralic. In B. Collinder (Ed.), *Comparative grammar of the Uralic languages* (pp. 250–251). Stockholm: Almqvist and Wiksell.

Sauerland, U. (2012). Experimental evidence for complex syntax in Pirahã. Accessed July, 2012 from http://ling.auf.net/lingBuzz/001095.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423, 623–656.

Tomalin, M. (2011). Syntactic structures and recursive devices: A legacy of imprecision. *Journal of Logic, Language, and Information*, *20*, 297–315.

Watumull, J., Hauser, M. D., Roberts, I. G., & Hornstein, N. (2014). On recursion. *Frontiers in Psychology*, *4*, 1–7.

Zséder, A., Recski, G., Varga, D., & Kornai, A. (2012). Rapid creation of large-scale corpora and frequency dictionaries. In *Proceedings to LREC 2012*, pp. 1462–1465.