

Chapter 8

Hierarchical syntactic structure in Malay/Indonesian, between Pirahã and Had Gadya

David Gil

Max Planck Institute for Evolutionary Anthropology

This paper presents an exploration of hierarchical syntactic structure in Malay/Indonesian. An analysis by Jackendoff & Wittenberg suggests that the Riau dialect of Indonesian may, like Pirahã, lack syntactic recursion. This paper poses the question whether Riau Indonesian is syntactically recursive, and answers “yes and no”. In Riau Indonesian, the grammar does indeed permit the formation of sentences with arbitrarily many levels of embedding; in this respect, it appears to exhibit syntactic recursion. However, in Riau Indonesian, the use of embedding is vanishingly rare: in a preliminary study involving comparable language registers, the frequency of embedding is found to be several orders of magnitude less in Malay/Indonesian than in English. To represent the status of recursion in Malay/Indonesian, reference must be made not only to grammar but also to the ways in which speakers choose to express their conceptualization of reality: linguistic perspective. A rough and ready indication of the propensity for a language to make use of syntactic recursion is provided by an examination of cumulative tales such as the Aramaic/Hebrew Passover song Had Gadya. Such tales occur throughout the world; however, while their semantics is recursive, their syntax is often flat and concatenative, a preliminary survey suggesting that the distribution of syntactically recursive cumulative tales may be limited to a single area centered on Europe and the Middle East. The results of this paper suggest that syntactic recursion is more likely to be found in languages spoken by communities of greater societal complexity, with Malay/Indonesian occupying an intermediate position in this respect, between Pirahã and Standard Average European.



1 Introduction

Dan Everett's main claim to fame, among linguists at least, is describing a language, Pirahã, arguably lacking many of the supposedly necessary design features of contemporary human language, such as numerals and quantifiers, colour terms, reference to things and events outside immediate experience, and of course, the cherry on the top, syntactic recursion (Everett 2005). Meanwhile, on the other side of the world, far from the spotlight, I have worked on describing the Riau dialect of Indonesian, which turns out to be lacking in a rather different set of core features, including clearly-defined words, distinct open syntactic categories, systematic encoding of thematic roles, and a grammaticalized distinction between attribution and predication (Gil 2005, 2006, 2012, 2013, 2017, 2020). This festschrift to Dan Everett presents an ideal opportunity to indulge in a contrastive analysis of the two languages – to see what happens when Pirahã meets Riau Indonesian. With neither the time nor the space to deal with the two languages in their entirety, I shall cut to the chase and focus on that most renowned feature associated with Dan Everett's work, namely syntactic recursion.

2 Is Riau Indonesian syntactically recursive?

In earlier conversations, Dan Everett (p.c.) suggested that Riau Indonesian may also be a language lacking in syntactic recursion. However, in his writings, for example Futrell et al. (2016), Everett (2017), he makes it clear that he bases this position on the analyses proposed by Jackendoff & Wittenberg (2014, 2017), who characterize Riau Indonesian as instantiating their class of Multi-Word Phrase Grammars, in which words group together to form phrases, and phrases group together to form utterances, but without any possibilities for syntactic recursion. Since Jackendoff & Wittenberg (2014, 2017) themselves base their analysis on my own descriptions, let us now take a look at what I have written on this score, and also some additional facts that I have not yet had the opportunity to describe in print. Spoiler alert: The answer to the question “Is Riau Indonesian syntactically recursive?” is “yes and no”.

Indeed, in Riau Indonesian, the grammar permits the formation of sentences such as the following:

- (1) Ali pukul orang yang suka anjing yang kejar anak.
Ali hit person PRTC like dog PRTC chase child
'Ali hit the person who likes the dog that chased the child.'

Moreover, if one really wants to, one can go on adding relative clauses at the end of the construction indefinitely. So in this respect, Riau Indonesian is clearly syntactically recursive, like English, not like Pirahã.

In previous work, my focus was largely on the grammar of simple word combinations, such as *ayam makan* (chicken eat), and thus, while making it clear that combinations of words can be grouped together to form ever larger combinations, this particular aspect of Riau Indonesian was not emphasized. This, then, is the appropriate time to redress the balance, by presenting an explicit analysis of hierarchical structure in Riau Indonesian. A syntactic analysis of sentence (1), following the theoretical framework laid out in Gil (2000), is presented in Figure 1.

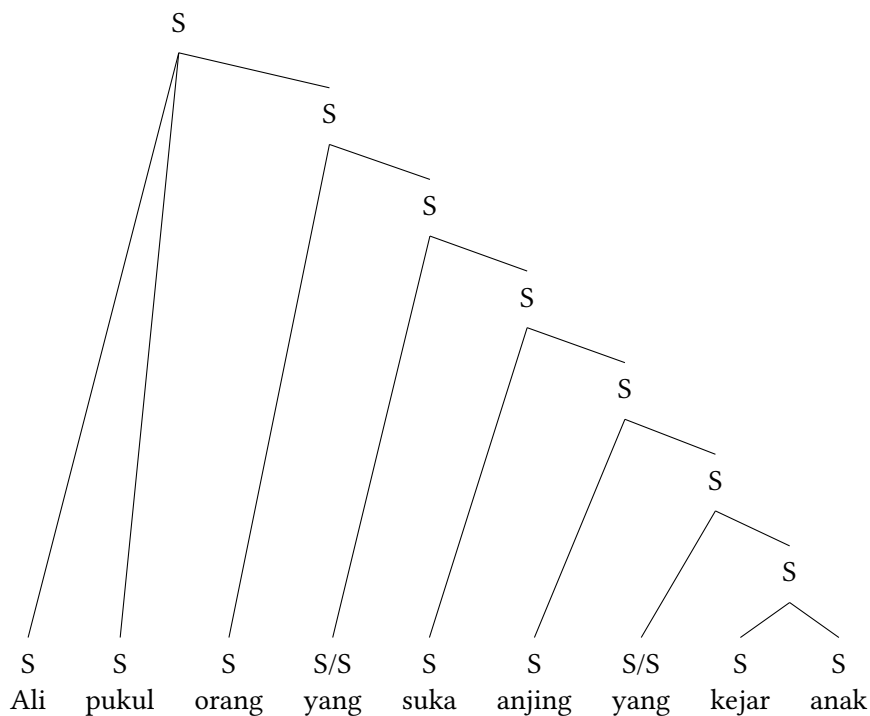


Figure 1: Syntactic structure of (1)

In Riau Indonesian, almost all words belong to the single open syntactic category S, for Sentence; this includes, among others, words denoting things, such as *anak* 'child', as well as words denoting activities, such as *kejar* 'chase'. The combination of the two, as in *kejar anak* above, is thus an instance of sentential

coordination. Similarly, complex Ss may combine with other simple or complex Ss to yield hierarchical structures such as that in Figure 1. While most words belong to the category S, a handful of grammatical items belong to a closed class, S/S, whose name, following the conventions of categorial grammar, specifies that they cannot occur on their own but rather must combine with Ss to yield superordinate Ss. Figure 1 contains two occurrences of the S/S word *yang*; in one case it combines with the S expression *kejar anak* to yield a superordinate S expression *yang kejar anak*, while in the other case it takes the S expression *suka anjing yang kejar anak* to yield a superordinate S expression *yang suka anjing yang kejar anak*. As suggested by the above, the syntax of Riau Indonesian is thus fully recursive, allowing for hierarchical syntactic structures of arbitrary depth.

As shown in Figure 2 below, the basic compositional semantics of sentence (1) makes reference to a hierarchical structure that is completely isomorphic to that of its syntax.

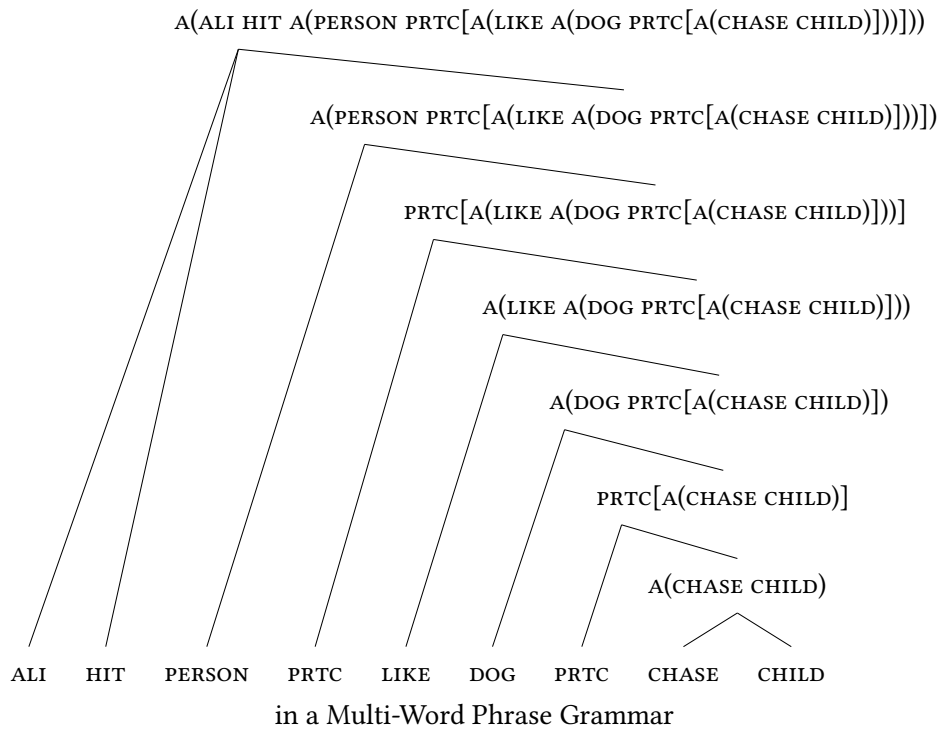


Figure 2: Structure of (1)

As represented in Figure 2, the semantics of sentence (1) makes reference to two operators. First and foremost is the association operator A, which underlies the

overwhelming majority of the compositional semantics. In its polyadic guise, illustrated above, it is a generalization of the monadic association operator, familiar as a semantic representation for genitive markers and various other possessive and associative constructions in many languages. Applying to two meanings, x and y , $A(x, y)$ simply means ‘entity associated with x and y ’, or ‘something to do with x and y ’. For example, in the above representation, $A(\text{CHASE CHILD})$, the meaning of *kejar anak*, means ‘entity associated with chase and child’; among many other things, it is unspecified for a variety of semantic categories such as number, definiteness, tense-aspect, and thematic role assignment (e.g. whether the child is the agent or patient of the chasing). The second operator is the participant operator PRTC , which, when applied to a meaning x , creates a new meaning $\text{PRTC}(x)$, picking out a participant in the semantic frame of x . The participant operator underlies the semantic representation of grammatical markers in several languages, which, under alternative analyses, are sometimes characterized as relativizers, nominalizers or reifiers. For example, in the above representation, $\text{PRTC}[A(\text{CHASE CHILD})]$ denotes a participant associated with the entity associated with *chase* and *child*, without any further specification of thematic role (i.e. whether it is the agent, the patient, or perhaps some other thematic role). The above analysis captures the fact that sentence (1) is massively underspecified with respect to categories such as number, definiteness, tense-aspect and thematic roles, with respect to which the English translation in (1), ‘Ali hit the person who likes the dog that chased the child’, is just one of myriad alternative available translations.

The above analysis may be contrasted with the alternative approach proposed by Jackendoff & Wittenberg (2014, 2017). One feature shared by both approaches is monocategoriality, the claim that Riau Indonesian has but a single open syntactic category. However, there are at least two significant differences between the two approaches. One pertains to the semantics, with respect to which Jackendoff & Wittenberg (2014, 2017) propose a more conventional analysis, based not on the association operator but rather on the familiar predicate-argument relationship. However, it is the second difference between the two approaches, pertaining to the syntax, that is of relevance to us here. In contrast to the above analysis, Jackendoff & Wittenberg (2014, 2017) assign Riau Indonesian to their class of Multi-Word Phrase Grammars, whose structures are flatter and non-recursive, making reference to just three categories, $W(\text{ord})$, $P(\text{hrase})$ and $U(\text{tterance})$. In fact, in their view, even these categories are more appropriately viewed as phonological, or prosodic, rather than syntactic. An analysis of sentence (1) in accordance with Jackendoff & Wittenberg (2014, 2017), whose approach is presented in Figure 3.

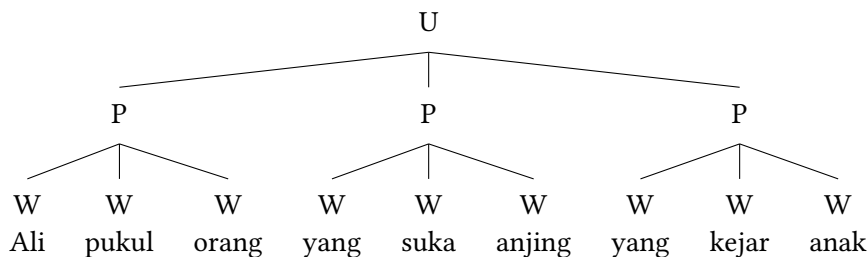


Figure 3: Semantic structure of (1)

In accordance with the analysis in Figure 3, sentence (1) does not exhibit multiple layers of hierarchical structure; instead it consists of a flat sequence of phrases, each consisting of a flat sequence of words. For Jackendoff & Wittenberg (2014, 2017), the obviously hierarchical nature of sentence (1) is a fact about its semantics, not its syntax.

How might one adjudicate between the alternative analyses of sentence (1) as represented in Figures 1 and 3 respectively? Given the flexibility of Riau Indonesian syntax as described in my earlier publications, Jackendoff & Wittenberg had good *prima facie* reason to invoke Occam's Razor and assign Riau Indonesian to their class of Multi-Word Phrase Grammars, entailing analyses such as that represented in Figure 3. Thus, with reference to Riau Indonesian, Jackendoff & Wittenberg (p.c.) argue that:

its grammar relates the semantic structure of sentences directly to linear order and prosodic constituency within phonology, without the intervention of syntax. This sort of grammar relies on principles such as Behaghel's First Law [...]. Agent precedes Patient, and Modifier precedes (or follows) Modified. None of these requires syntax.

Nevertheless, additional evidence suggests that the flat structures of Multi-Word Phrase Grammars are insufficient to account for the totality of facts pertaining to the syntax of Riau Indonesian.

As alluded to in the above passage by Jackendoff & Wittenberg (p.c.), one of the core characteristics of Riau Indonesian sentence structure is the extent, greater than in many other languages, to which it upholds Behaghel's First Law, which, paraphrasing slightly, states that expressions whose meanings are closer to each other in conceptual space tend to occur closer to each other within the syntactic structure of the sentence (Behaghel 1932). One manifestation of the strongly

Behaghelian nature of Riau Indonesian is the relative infrequency with which expressions occur outside of their “expected” positions – the kinds of constructions that, within some theoretical frameworks, are often accounted for in terms of various kinds of movement rules. In the case at hand, Behaghel’s First Law provides clear-cut and unambivalent motivation for several of the intermediate levels of syntactic structure posited by the analysis in Figure 1 but absent from the flatter structure posited by the analysis in Figure 3.

Consider, for example, the substring of words *anjing yang kejar anak* (dog PRTC chase child) in (1), which forms a constituent in Figure 1 but not in Figure 3. This constituency reflects the fact that these four words are close to each other semantically, *anjing* and *anak* being participants in the activity denoted by *kejar*. In doing so, it makes correct predictions about possible reorderings of the words in (1). For example, it correctly predicts that *anjing yang kejar anak* can be moved, as a single chunk, to the beginning of the sentence, as in (2) below:

- (2) *Anjing yang kejar anak Ali pukul orang yang suka.*
 dog PRTC chase child Ali hit person PRTC like
 ‘The dog that chased the child, Ali hit the person who likes it.’

Conversely, it also correctly predicts that *yang suka anjing*, a constituent in Figure 3 but not in Figure 1, cannot be moved to the beginning of the sentence, as in (3) below, without a massive change in meaning:

- (3) *Yang suka anjing Ali pukul orang yang kejar anak.*
 PRTC like dog Ali hit person PRTC chase child
 ‘As for the one who likes the dog, Ali hit the person who chased the child.’
 * ‘Ali hit the person who likes the dog that chased the child.’ [= (1)]

As indicated above, the interpretation associated with sentence (1) is completely unavailable in (3). More dramatically, the hierarchical syntactic structure of Figure 1 predicts the total unacceptability, *salva veritate*, of various random reshufflings of the words in (1) such as in (4) below:

- (4) *Orang pukul anjing yang suka anak yang kejar Ali.*
 person hit dog PRTC like child PRTC chase Ali
 ‘A person hit the dog that likes the child who chased Ali.’
 * ‘Ali hit the person who likes the dog that chased the child.’ [= (1)]

Again, the interpretation associated with sentence (1) is completely unavailable in (4). Thus, as shown above, Riau Indonesian has nothing of the freedom of, say,

a language like Warlpiri, for which Hale (1979, 1983) posits a flat “W-star” grammar. Instead, the strongly Behagelian nature of Riau Indonesian, compensating for its flexibility and indeterminacy in various other domains, provides strong support for hierarchical syntactic structures of the kind represented in Figure 1, and in doing so for the presence of syntactic recursion in Riau Indonesian.

Jackendoff & Wittenberg (p.c.) propose to account for facts similar to these by positing a “Contiguity Condition”, whose effect is to ensure that what they call “semantic constituents” must be expressed by contiguous strings of words. For example, in order to account for the inseparability of the string *anjing yang kejar anak*, as evidenced in (1–4) above, they would invoke the Contiguity Condition to ensure that the semantic constituent A(DOG PRTC[A(CHASE CHILD)]) is realized by a contiguous string of words. Crucially, according to Jackendoff & Wittenberg, such strings do not form syntactic constituents; rather, they are semantic constituents that are then mapped onto phonological ones. Their analysis is in accordance with their general agenda, which is to avoid reference to syntax unless the phenomenon under question cannot be accounted for with reference to either semantics or phonology. While in general I am supportive of their agenda, in the case at hand there is an obvious problem, namely, their assumption that contiguity is a fact about phonology. While in some cases, indeed, contiguous words may interact phonologically, for example by forming an intonational phrase, this is not true more generally; for example, in (1), the string *anjing yang kejar anak*, argued in (2) to form a syntactic constituent, does not constitute a complete intonational phrase. Rather, contiguity must be viewed as a syntactic property, indeed perhaps the most quintessential one, pertaining to what syntax is all about, namely, the bringing together of expressions in order to constitute larger expressions. Thus, pace Jackendoff & Wittenberg, the strongly Behagelian nature of Riau Indonesian, as exemplified by the above examples, does indeed support the kind of hierarchical syntactic structure illustrated in Figure 1, and in so doing also the presence of syntactic recursion in Riau Indonesian.

With respect to the presence of syntactic recursion, then, Riau Indonesian is more like English than like Pirahã. But still, this is only half the story. A more fully adequate account of recursion in Riau Indonesian must acknowledge the fact that sentences such as in (1) are highly artificial, and that nobody actually speaks that way. Exactly what it means to say that nobody speaks that way is taken up in the next section.

3 Linguistic perspective (or, what it means to know a language)

Among the scholars who have addressed the broader implications of my analysis of Riau Indonesian is James Hurford. In Hurford (2011: 410–413), an email conversation is reproduced, in which the author asks me various questions about the apparent extreme simplicity of Riau Indonesian and its implications for the evolution of language. Towards the end of the conversation, the following exchange takes place:

Hurford: [If Riau Indonesian grammar is as simple as you describe,] what is there to learn, beside vocabulary? How come you need a full-time teacher?

Gil: The grammar, in the narrow Chomskyan sense of ‘set of well-formed strings’, can be learned in less than an hour. But still, in order to be able to be mistaken for a native speaker down a dark alley, you’d need to spend years learning: lexicon, phonetics, and, most interestingly, that nebulous domain that is sometimes referred to as idiomaticity – being able to say something that is not just grammatical but also stylistically felicitous in the appropriate context.

Hurford: I’m pondering what you mean by “idiomaticity” ...

Hurford’s trouble with the term “idiomaticity” is understandable, as I was not very clear back then with regard to what I meant by this particular term. This, then, is the appropriate place to try and clarify matters.

The notion in question is one that has been put forward, in various guises, by a number of different scholars. Grace (1987) talks of the “linguistic construction of reality”, Pawley (1993) – from whom I adapted the above term – refers to “idiomatic competence”, Slobin (1996: 91) alludes to a “subjective orientation of the world of experience”, while Ross (2001) discusses the different “ways of saying things” associated with different languages. In this paper I propose the term “linguistic perspective” – see Gil (2023) for a detailed application of this notion to the field of diachronic syntax and language contact.

The leading idea is as follows. The reality in which we find ourselves is of overwhelming complexity, as also is our conceptualization of it, which is what we express by means of language. However, any natural human language can only express a small proportion of this complexity. Hence, using language involves choosing which aspects of our conceptualization of reality are worthy of expression, and which others are to be left unexpressed. Or, in other words, adopting a

perspective. While in some instances, such choices involve on-the-fly decisions by individual speakers, in many other cases, these choices are conventionalized, at the level of the language and the community of which it forms part. And of course, different languages and the communities in which they are embedded conventionalize different choices, as a result of which they may be said to be associated with different linguistic perspectives.

Some of the differences in linguistic perspective between Riau Indonesian and English may be illustrated with reference to an everyday situation, in which you are walking down the street with your friend and notice that he has just dropped his wallet. What might you say? Here are two natural utterances in Riau Indonesian and English respectively:

- (5) Dompét jatuh bang.
wallet fall HYP/elder.brother
- (6) You dropped your wallet.

While the situation is the same, the two languages choose to express different aspects of it; they adopt different perspectives. In Riau Indonesian, interpersonal relationships feature prominently, and are typically expressed by kinship terms, such as the hypocoristic form *bang* in (5). In contrast, in English, concepts such as time and number are commonly encoded, as is exemplified in (6) by the past-tense *-ed* suffix on *drop* and the absence of a plural marker on *wallet*; in addition, the notions of participant and possession are also expressed by means of the pronominal forms *you* and *your*. Although the expression of these different perspectives makes use of lexicon and grammar, the perspectives themselves are not part of the lexicon or the grammar, but rather a completely separate component of the language, which speakers have to master in order to be able to speak the language properly.

The independence of linguistic perspective from lexicon and grammar can be seen most clearly by consideration of the following variants of (5) and (6), in which the utterances are couched in the perspective of the “wrong” language:

- (7) # Kamu tadi jatuh dompet kamu satu.
2 PST.PROX fall wallet 2 one
- (8) # Wallet drop bro.

The strangeness of examples (7) and (8) is indicated above by the use of the symbol #. Sentence (7) in Riau Indonesian is as precise a rendition as is possible of sentence (6) in English, leaving out the term of address, but including expression of time, with *tadi*; number, with the numeral *satu*, participant, with the

pronoun *kamu*, and possession, with the second occurrence of *kamu* in a possessive construction. Sentence (7) is perfectly grammatical in Riau Indonesian, but is hopelessly long-winded; nobody would ever say anything like that in real life. Conversely, sentence (8) in English is perhaps as close as one can get to sentence (5) in Riau Indonesian, leaving out the expression of time, number, participant and possession, but including instead the kinship term *bro*. While the bare verb *drop* is of marginal grammaticality in the given syntactic environment, a further major problem in (8) involves the use of inappropriate and clashing registers: whereas the bare noun phrase *wallet* is associated with the telegraphic language of newspaper headlines and the like, the address term *bro* is restricted in its occurrence to certain speech styles of particular subcultures. In summary, what is strange about sentence (7) and in large part also (8) is a matter of linguistic perspective, not lexicon or grammar.

The notion of linguistic perspective provides the basis for a proper understanding of how Riau Indonesian can be syntactically recursive in one respect but not in another. Specifically, sentence (1) is strange in a similar way to sentences (7) and (8) above; it should also have been marked with a #. While illustrating the recursive potential of Riau Indonesian syntax, it violates the linguistic perspective of Riau Indonesian. Specifically, in Riau Indonesian, speakers systematically choose not to afford overt morphosyntactic expression to the kind of logical subordination that is commonly expressed in English by multiple syntactic embedding. Thus, in a situation that might prompt a speaker of English to say something like 'Ali hit the person who likes the dog that chased the child', a speaker of Riau Indonesian is much less likely to produce a sentence such as (1), and instead more likely to produce one that might broadly resemble the following:

- (9) Ada anak kan, dia kejar anjing, terus ada orang suka anjing
 exist child Q 3 chase dog continue exist person like dog
 tu, orangnya kena pukul Ali.
 DEM.PROX person:ASSOC undergo hit Ali
 'There was a child, right, he was chased by a dog, then there was a man
 who liked the dog, and the man got hit by Ali'.

Sentence (9) above consists of four clauses strung out one after the other in a loose paratactic relationship, in which the logical subordination expressed in English, as well as in the bizarre Riau Indonesian sentence (1), by means of syntactic embedding, is here manifest mostly by relationships of coreference between pairs of expressions, namely *anak* and *dia*, *anjing* and *anjing tu*, and *orang* and *orangnya*, with just a single instance of embedding, in which the expression *suka*

anjing tu is subordinate to *ada orang*. Thus, the reluctance of Riau Indonesian speakers to make use of the grammatical devices available to express relationships of multiple logical subordination may be viewed as another aspect of the linguistic perspective of Riau Indonesian, determining which aspects of reality are linguistically encoded and which others are left unexpressed.

The contrasting linguistic perspectives of Riau Indonesian and English may be represented diagrammatically as in Figure 4.

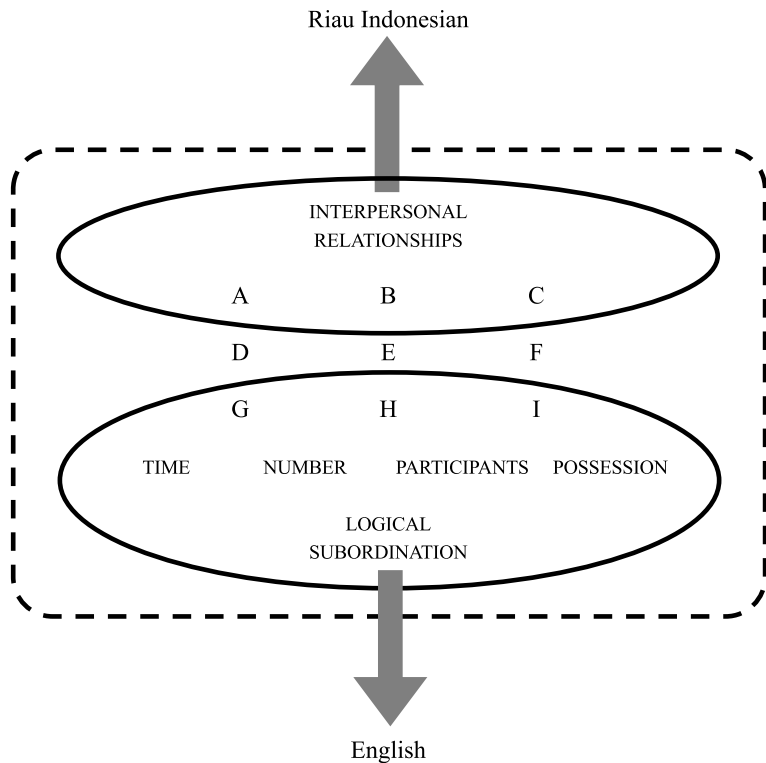


Figure 4: Linguistic perspective in Riau Indonesian and English

In Figure 4, the area enclosed by the dashed line represents our conceptualization of reality. Within it, represented in small caps, are a variety of individual aspects of our conceptualization of reality which may potentially be expressed in language. While some of these aspects are the ones discussed above, a potentially much larger number of other aspects are suggested by the letters A–I. The two ovals show how Riau Indonesian and English select different subsets of these variegated aspects of reality; they thus represent the contrasting linguistic perspectives of the two languages. In particular, they capture the fact that even

though both languages possess formal devices for the expression of logical subordination, actually using them to form multiple embeddings is something that speakers of English often do while speakers of Riau Indonesian typically do not.

The preceding discussion thus provides an answer to Jim Hurford's question posed at the beginning of this section, namely, what was meant by "idiomaticity". In doing so, it underscores the fact that knowing a language involves mastery not only of lexicon, phonetics, phonology, morphology, syntax, semantics and discourse/pragmatics, but, crucially, also linguistic perspective. In particular, in order to be able to speak Riau Indonesian properly, it is not enough to know how to construct multiple embeddings; in addition, one must also know that the actual use of such multiple embeddings is, in almost all cases, inappropriate – it's just not the Riau Indonesian way of speaking.

4 Quantitative comparison

It may reasonably be objected that it is unfair to compare a standardized language such as English with a colloquial variety such as Riau Indonesian. After all, there is substantial evidence that within many a language, subordination occurs more frequently in higher registers than in lower ones, and in written as opposed to oral modalities (Givón 1979, Deutscher 2000, Karlsson 2009a,b, and others). Indeed, corpora of spoken English are also likely to contain constructions similar to the English rendition of (9), 'There was a child, right, he was chased by a dog, then there was a man who liked the dog, and the man got hit by Ali'.

Nevertheless, abstracting away from such language-internal variation, English as a whole is much more tolerant of multiple embedding than Malay/Indonesian, the macro-language of which Riau Indonesian is just one typical exemplar. In order to support this claim empirically, a rough and ready corpus study was conducted making use of Google searches. In both English and Malay/Indonesian, Google searches produce material representative not only of a range of different dialects, but also of a variety of registers ranging from the colloquial language of social media to the more formal language of written works. Thus, it is reasonable to believe that comparing the results of Google searches in English and Malay/Indonesian abstracts away not only from modality effects but also from the effects of language-internal register-governed variation; such searches may thus potentially offer a fair and well-founded comparison of the two languages.

Some results of a comparative corpus study of English and Malay/Indonesian are presented in Tables 1–4 below.

Tables 1–4 present two comparisons of English and Malay/Indonesian, the first in Tables 1 and 2 and the second in Tables 3 and 4. In Tables 1–4, the first column

Table 1: First comparison: English

he thinks he said	13,840,000	
he	4,870,000,000	7×10^{-4}
thinks	4,260,000,000	9×10^{-4}
said	4,560,000,000	8×10^{-4}

Table 2: First comparison: Malay/Indonesian

dia pikir dia bilang	7	
dia (3sg)	4,060,000,000	2×10^{-9}
pikir (think)	38,000,000	2×10^{-7}
bilang (say)	1,210,000,000	6×10^{-9}

Table 3: Second comparison: English

he said he thinks	10,200,000	
he	4,870,000,000	2×10^{-3}
thinks	4,260,000,000	2×10^{-3}
said	4,560,000,000	2×10^{-3}

Table 4: Second comparison: Malay/Indonesian

dia bilang dia pikir	37,800	
dia (3sg)	4,060,000,000	9×10^{-6}
pikir (think)	38,000,000	1×10^{-3}
bilang (say)	1,210,000,000	3×10^{-5}

presents the search criterion, and the second column the rough number of hits (accessed on the 5th of June 2022). Within each table, the first row presents a string, part of a complex construction which, when adding some following text, involves two levels of embedding, while the subsequent rows present the individual words occurring within the complex construction.

By examining the ratio of hits for the complex construction in the first line to that of the individual words in the subsequent lines, it is possible to abstract away both from the different sizes of the two corpora and also from possible frequency effects associated with the individual words, and in so doing measure the propensity of the language to form multiple embeddings making use of the words in question. These ratios are presented in rounded form in the third column of each table. For example, in Table 1, the ratio of *he* in line 2 to *he thinks he said* in line 1 is $3,840,000/4,870,000,000 = 7 \times 10^{-4}$. On its own, this figure does not mean much; what is significant is its comparison to the corresponding figure in Table 2, in which the the ratio of *dia* in line 2 to *dia pikir dia bilang* in line 1 is $7/4,060,000,000 = 2 \times 10^{-9}$. Comparing these two figures, 7×10^{-4} and 2×10^{-9} , shows that with respect to the particular words chosen, English is 4 to 5 orders of magnitude more likely to form the multiple embedding construction than Malay/Indonesian. Comparable order-of-magnitude differences are present for the remaining five comparisons between Tables 1 and 2, and for five out of six of the corresponding comparisons between Tables 3 and 4 – the only case where there is not a order-of-magnitude difference being in the third lines of Tables 3 and 4, where the English is “only” about twice as likely to form a multiple embedding construction than the Malay/Indonesian.

Thus, Tables 1–4 show that English as a whole is massively more likely to form multiple embedding constructions than Malay/Indonesian. In yet another comparison aimed at testing this generalization, the English string *who do you think is going to win* yielded a total of 841,000 hits, while there were no hits whatsoever for numerous potential equivalent sentences in Malay/Indonesian, the only exception being 3 hits for *siapa anda pikir akan menang* (who 2 think FUT win), which actually occurred in an Indonesian website explaining the meaning of a similar English sentence – the exception that proves the rule.

The above Google searches show that multiple embedding is massively more common in English than in Malay/Indonesian. The similar nature of the searches in the two languages shows that the greater propensity for subordination in English as compared to Malay/Indonesian is a cross-linguistic difference that is independent of both register and modality. In particular, the more widespread use of embedding in English is observable notwithstanding a significant body of literature (Karlsson 2007b,a, 2009a,b) showing that such constructions are more

highly constrained, and occur less frequently in real-life language use, than is commonly assumed to be the case under a simplistic characterization of English and other similar languages as syntactically recursive.

The results of the comparison between syntactic recursion in Malay/Indonesian and English is summarized in Figure 5.

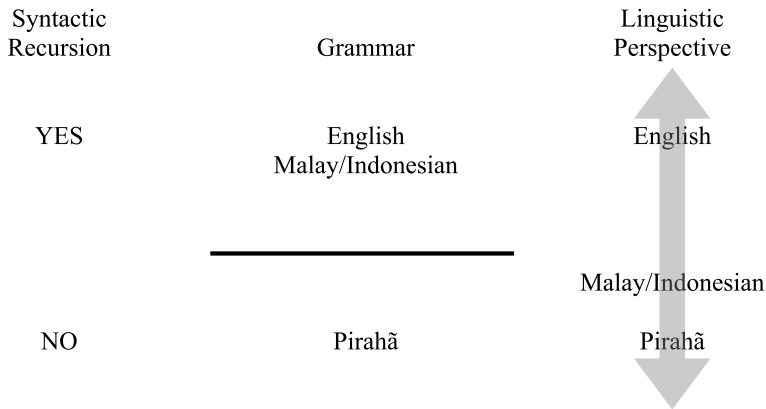


Figure 5: Recursion in Riau Indonesian and English

In terms of grammar, the distinction between having recursion and not having it is categorical; Malay/Indonesian (including Riau Indonesian) like English, has it, while Pirahã does not. In contrast, with respect to linguistic perspective, the propensity for syntactic subordination and the use of recursive strategies to effect such subordination is a scalar property; while English makes frequent use of such embedding, Malay/Indonesian does not, though perhaps – this remains an open question – not to quite the same extent as Pirahã. Thus, Figure 5 shows why the question posed earlier, whether Riau Indonesian has syntactic recursion, is appropriately answered with a “yes and no”. Answering the question in this way underscores the importance, in any description of a language, not only of its lexicon and grammar, the things that one can say, but also of its linguistic perspective, the things that one actually does say.

5 Towards a cross-linguistic typology

How do other languages fit into the above schema: are more languages like English, like Malay/Indonesian, or like Pirahã? In addition, since linguistic perspective offers a scalar rather than a categorical take on recursion, there is also the potential for languages to occupy other points on the scale, possibly higher

than English, or in-between English and Malay/Indonesian, with respect to their propensity for multiple embedding and their associated degrees of recursion.

At present, we simply do not know enough to provide a systematic answer to the above question. Recursion is in the eyes of the beholder, with different analyses pointing towards alternative conclusions with regard to whether a given construction constitutes an instantiation of subordination or not. If linguist A says that language X is syntactically recursive, while linguist B argues that language Y is not, then is this a real difference between the two languages, or merely a difference in the ways the two linguists choose to pursue their trade? Indeed, for many individual languages, different linguists offer different answers to the question whether the language is syntactically recursive, as we saw earlier with the contrasting analyses of Riau Indonesian offered by myself and by Jackendoff & Wittenberg. In fact, in some cases, it is not different linguists proposing different analyses but the same linguist modifying their views over time, an apparent instance of this being Dan Everett on Pirahã, as described in detail by Pullum (2024), Chapter 2 of this volume.

As reflected by the evolution of Everett's writings on Pirahã, whether or not linguists choose to analyze a particular language as exhibiting syntactic recursion may be affected in a systematic fashion by their methodology, and, in particular, the relative weights that they attribute to data deriving from elicitation as opposed to naturalistic corpora. In general, elicitation is more likely to lead to insights into grammar, and what people *can* say, whereas work based on naturalistic corpora stands a greater chance of yielding a better understanding of linguistic perspective, and what people actually *do* say. Thus, elicitation, and asking speakers whether they would be willing to accept a long and unwieldy construction, is more conducive to the positing of syntactic recursion, whereas observation of naturalistic texts, in which such constructions may occur rarely, if at all, is more likely to lead to a claim to the effect that syntactic recursion is absent. In particular, claims by scholars such as Sandalo et al. (2018), and indeed for that matter also the early Everett (1986), to the effect that the grammar of Pirahã has at least some syntactic subordination, tend to be the products of elicited data, motivated by theoretical concerns, and as shown in detail by Pullum (2024), Chapter 2 of this volume, a desire to fit the language into a predetermined descriptive template. In contrast, the later and more famous claims by Dan Everett that Pirahã lacks syntactic recursion are mostly based on the use of naturalistic data. While for Pirahã it may indeed be the case, as argued by Pullum (2024), Chapter 2 of this volume, that the work based on elicitation is of inferior quality to the work based on naturalistic data, this is a contingent fact and not an inherent property of elicitation as opposed to the use of naturalistic corpora – there

can be good or bad work based on elicitation just as there can be good or bad work making use of naturalistic data. Thus, in view of the diverse methodologies underlying the available descriptions and analyses of different languages, it is not yet possible to paint a systematic picture of cross-linguistic variation with regard to syntactic recursion, but only to offer some observations and conjectures.

To begin with, one may ask whether there are languages in which syntactic recursion is even more prevalent than in English. One phenomenon that comes to mind is that of clause chaining in many of the languages of New Guinea, as described by Foley (2010) and others. In such languages, a sequence of several monoclausal sentences in English are rendered in the form of a single complex clause, in which all but the last of the clauses constitute a chain of embedded clauses marked by various morphosyntactic devices as subordinate. However, as suggested by the term “chaining”, the non-final clauses are all of equal status to one another, and therefore do not exhibit the kind of multiple embedding that is of concern to us here.

A perhaps more promising case is argued by Cysouw (2023) to be provided by German, as illustrated by sentences such as the following (from *Juristische Schulung, Zeitschrift für Studium und Referendariat*, 10/2012, Verlag C.H. Beck, p. 866; the English translation was provided by Boban Arsenijević with the assistance of DeepL Translate.):

- (10) Solange sich der Gläubiger noch durch die nachgeholte Leistung in Natur, ggf. ergänzt durch den Ersatz von Verzögerungs- und sonstigen Schäden, vollständig in die Lage versetzen lässt, in der er sich bei pflichtgemäßer Leistung befände, und die Leistung für den Schuldner weniger kostspielig ist als die Zahlung von Schadensersatz statt der Leistung, gibt es keinen Grund, dem Gläubiger von vornherein die Entscheidung zwischen Erfüllung in Natur und Schadensersatz zu überlassen und ihm zu erlauben, dem Schuldner durch die Wahl des Schadensersatzes den Kostenvorteil der Leistungserbringung zu nehmen.

‘As long as the creditor can still be fully put in the position he would be in if he had performed dutifully, by the subsequent performance in nature, supplemented, if necessary, by the compensation for delay and other damages, and the performance is less costly for the debtor than the payment of damages instead of performance, there is no reason to leave the creditor the choice between performance in nature and damages from the outset and to allow him to deprive the debtor of the cost advantage by choosing the way of damage compensation.’

According to Cysouw (2023), such sentences, complex to the point of unintelligibility for the average reader, are more common in German than in English.

To the extent that this observation holds water, it would seem to indicate that German may be characterized by a greater propensity for multiple embedding than English. Some further discussion of stylistic variation with respect to the propensity for various kinds of subordination in the languages of Europe can be found in Karlsson (2007b,a), suggesting, *inter alia*, that contemporary English might represent the outcome of recent processes of reduction in the extent to which such constructions, involving multiple embedding, are used.

At the other end of the scale represented in Figure 5 are languages whose grammars do not allow syntactic recursion, or whose linguistic perspectives and associated patterns of usage disfavour it. Again, one may ask whether and to what extent Malay/Indonesian and Pirahã are weird outliers, and the answer, surprising perhaps only to those whose primary concern is English and its representation within certain contemporary syntactic theories, is that they are not at all exceptional. Thus, Givón (1979: 298) writes that “there are some languages extant to this day – all in preindustrial, illiterate societies with relatively small, homogenous social units – where one could demonstrate that subordination does not really exist ...”. Givón’s assertion is cited approvingly by Pullum & Scholz (2010) and Pullum (2024), Chapter 2 of this volume, who go on to adduce several descriptions of languages reported as lacking subordination, among which are Amazonian languages such as Hixkaryana (Derbyshire 1979) and Apalaí (Koehn & Koehn 1986), Australian languages such as Dyirbal (Dixon 1972) and Wargamay (Dixon 1981), and various ancient languages, either attested, such as Old Akkadian (Deutscher 2000), or reconstructed, such as Proto-Uralic (Collinder 1960). And doubtlessly there are many more such languages, whose existence may have been obscured by a common analytical bias that leads us to seek out complex structures where none actually exist. Thus, with respect to syntactic recursion, at least, Pirahã is in good company, and is anything but some kind of strange outlandish creature, or, as intimated by some of Dan Everett’s critics, something even worse than that. Indeed, given the large number of languages still associated with such “preindustrial, illiterate societies”, one is even led to wonder whether such languages might constitute the cross-linguistic norm.

And what of Riau Indonesian? As suggested above, its extreme disfavoured of subordination is quite unexceptional; this is true not only within Malay/Indonesian (see Englebretson 2003 for a similar take on another colloquial variety of Indonesian, spoken on the island of Java) but also cross-linguistically. Nevertheless, the availability, however dispreferred, of constructions such as that in (1) shows that Riau Indonesian is not quite like Pirahã or any of the other languages cited above. This rather extreme clash between what people can say and how people do actually speak is due, at least in part, to the much more complex

sociolinguistic circumstances associated with Riau Indonesian, and other similar colloquial varieties of Malay/Indonesian. Such colloquial varieties of Malay/Indonesian occupy the bottom reaches of a lectal cline, at whose upper end are the two standardized versions of the language, Standard Indonesian and Standard Malay. Although structurally quite distinct from one another, colloquial and standard versions of Malay/Indonesian are both present in the minds of diglossic speakers, and as a result, each of the two ends of the cline, colloquial and standard, exerts a substantial effect on its counterpart at the opposite end. On the one hand, it is the presence of multiple embedding in Standard Indonesian that allows a speaker of Riau Indonesian to accept, however reluctantly, constructions such as that in (1), thereby supporting the characterization of Riau Indonesian grammar as syntactically recursive. In this respect, then, Riau Indonesian differs from Pirahã and other similar languages lacking a standardized acrolect that might be more conducive to such recursion. On the other hand, it is the extreme disfavours of subordination in Riau and other colloquial varieties of Indonesian that percolates upwards along the lectal cline, resulting in a relative disfavours of subordination also in Standard Indonesian and Malay, as reflected by the differential results of the Google searches reported on in Section 4 earlier. In this regard, Malay/Indonesian presents a clear contrast to English. Standard Malay and Indonesian constitute special registers, not acquired natively by speakers through the usual processes of first language acquisition; they are thus parasitic on their colloquial counterparts (Gil, to appear). In contrast, Standard English is a more natural language variety that is indeed acquired natively by a large population of speakers, and is therefore relatively more resistant to influences from basilectal varieties of English, as, for example, might be manifest in the disfavours of subordination.

The existence of cross-linguistic variation with respect to syntactic recursion and the proclivity for subordination raises the question what the determining factors might be that underlie such variation. Givón, in the above-cited passage, alludes to “preindustrial, illiterate societies with relatively small, homogenous social units” – a plausible hypothesis, but one still in need of solid empirical cross-linguistic support. As suggested earlier, a major challenge faced by any attempt to seek such support is that whether or not a language has syntactic recursion is very much dependent on how it is analyzed. What is needed, therefore, is a common yardstick by which different languages can be uniformly and objectively assessed with respect to their relative proclivities for syntactic subordination.

6 Cumulative tales

A simple rough and ready measure for the assessment of syntactic recursion across languages is provided by the analysis of an easily-identifiable genre of verbal art, namely cumulative tales (Thompson 1946: 230–234, Aarne & Thompson 1961: 522–536).

A well-known example of a cumulative tale is the Jewish Passover song *Had Gadya*, mostly in Aramaic though with some Hebrew interspersed, shown below transcribed in accordance with a Modern Hebrew pronunciation, followed by an English translation (by Eve Levavi, in *Haggadah for Pesah, an English translation*, hosted on the Open Siddur Project):

- (11) Veata hakadoš barux hu vešaḥat lemalʔax hamavet,
 dešaḥat lešoḥet,
 dešaḥat letora,
 dešata lemaya,
 dexava lenura,
 desaraf leḥutra,
 dehika lekalba,
 denašax lešunra,
 deaxla legadya,
 dezabin aba bitrey zuzey.
 ‘Then the Holy One, Blessed be He, came and slaughtered the angel of
 death,
 who slaughtered the butcher,
 who slaughtered the ox,
 that drank the water,
 that put out the fire,
 that burned the stick,
 that beat the dog,
 that bit the cat,
 that ate the goat,
 that my father bought for two zuzim.’

In (11) above, the last verse is presented, containing a total of 10 clauses, of which the last 9, introduced by the Aramaic relativizer *de-*, are embedded, each within the clause immediately preceding it, like a set of Matryoshka dolls.

The clear and well-defined structural properties of cumulative tales provide an objective criterion enabling their cross-linguistic and cross-cultural distribution

to be gauged. In the absence – as far as I was able to determine – of any existing systematic cross-linguistic study of cumulative tales, a query was posted on the LINGTYP list¹ in which its diverse readership, encompassing typologists familiar with a wide range of the world's languages, were asked if they were familiar with examples of cumulative tales or songs from their respective regions of expertise. The responses that came in demonstrated that cumulative tales are indeed widespread across the world's languages and cultures, but with a crucial qualification. It turns out that the Had Gadya type, involving syntactic recursion, is apparently the exception, and that in most cases, the recursion is of a purely semantic nature, and is not reflected by syntactic subordination – the formal relationship between the successive clauses instead being one of parataxis, or flat chaining. An instance of the latter state of affairs is provided by the following example from an Alemannic dialect of German (cited in Meier 1851: 65 and Newell 1905: 39, and translated into English by Claudia Wegener):

- (12) Gestern haun i fegelt, I haun e Kreuzer gwonne;
De Kreuzer haun u 'r Mutterb gean,
Mutter hat mir Kerne gean,
Kerne haun i 'm Müller gean,
Müller hat mir Mehl gean,
Mehl haun i 'm Becke gean,
Becke hat mir Wede gean,
Wede haun i 'm Vater gean,
Vater hat mir e Stöckle gean,
Stöckle haun i 'm Lehrer hean.
Lehrer hat mir Tatze hean ...
'Yesterday I went bowling, I won a kreutzer;
The kreutzer I gave to my mother,
My mother gave me corn,
The corn I gave to the miller,
The miller gave me flour,
The flour I gave to the baker,
The baker gave me a bun,
The bun I gave to my father,
My father gave me a stick,
The stick I gave to my teacher,
My teacher hit my hand ...'

¹<https://listserv.linguistlist.org/pipermail/lingtyp/2023-February/thread.html>

Although in terms of its semantic structure, example (12) bears a close resemblance to its predecessor in (11), in the Allemanic case there is no syntactic subordination, but instead a sequence of independent clauses, connected to each other by successive relationships of coreferentiality between the subject NP of each clause and the object NP of the clause preceding it. Syntactically, then, rather than Matryoshka dolls, the clauses in (12) are like beads on a string.

While the results of the LINGTYP survey cannot be considered more than preliminary, the emerging picture is one in which cumulative tales involving parataxis, as in example (12), are of widespread distribution, occurring in a variety of unrelated languages and cultural regions, and hence not accountable for in terms of a single ancestral case subsequently spreading by means of vertical inheritance or horizontal diffusion. Some of the languages in which such syntactically flat cumulative tales are attested include Laal, an isolate language of southern Chad (Florian Lionnet p.c.); the Timimoun dialect of Arabic (Mammeri 1985, Lameen Souag p.c.), Basque (Peter Bakker p.c.), the Nakh-Daghestanian language Agul (Timor Maisak p.c.), Tamil (Siva Kalyan p.c.), three at best distantly related languages of New Guinea, Pa, Northeastern Kiwai and Asmat (Voorhoeve 2010), and Yucatec Maya (Smailus 1975: 180–186, Jürgen Bohnemeyer p.c.). In contrast, no examples emerged of syntactically recursive cumulative tales, as in (11), from outside the Middle-Eastern/North African/European cultural area – even though, in many such languages, syntactic recursion is permitted by the grammar, and at least some syntactic subordination is used in ordinary speech.

If indeed the limited distribution of syntactically recursive cumulative tales is found to hold up under more extensive empirical scrutiny, then this would potentially point towards one or both of the following two hypotheses governing their distribution. The first hypothesis is diachronic, suggesting that in accordance with this hypothesis, all cases of syntactically recursive cumulative tales can be traced back to a single common origin, with their current distribution the outcome of an interplay of vertical inheritance and horizontal diffusion. While it is beyond doubt that literary genres such as cumulative tales do indeed spread across time and space, what is at issue is whether such diachronic processes are the whole story, or whether other factors might also play a role in the observed distribution of syntactically recursive cumulative tales.

The second hypothesis offers one such factor of a sociolinguistic orientation, appealing to the covariance of societal and grammatical complexity. Specifically, the occurrence of syntactically recursive cumulative tales correlates positively with the complexity of the society with which it is associated, as measured by any of a number of potential yardsticks, in accordance with Chen et al. (2024). In alternative formulations of the hypothesis, syntactically recursive cumulative

tales would be more likely to occur in exoteric as opposed to esoteric societies (Thurston 1987, Wray & Grace 2007); societies characterized as Western, Educated, Industrialized, Rich and Democratic, or WEIRD (Henrich et al. 2010, Henrich 2020) or as Literate, Official, and with Lots of speakers, or LOL (Dahl 2015); societies high on Ethnologue's Expanded Graded Intergenerational Disruption Scale, or EGIDS (Eberhard et al. 2020); and societies of greater complexity as reflected in a variety of features in the D-PLACE database (Kirby et al. 2016), such as, for example, population size, population density, and the number of jurisdictional levels above the local community. While the results of the informal LINGTYP survey are consistent with this hypothesis, more work is necessary before such a hypothesis can be empirically supported.

7 Conclusion

However interesting in its own right, the examination of the distribution of syntactically recursive cumulative tales is of course intended here as a mere proxy for a more general investigation into the cross-linguistic distribution of syntactic recursion. While much harder to ascertain in an objective systematic manner, there is good reason to believe that the distribution of syntactic recursion more generally may be sensitive to the variegated aspects of societal complexity alluded to in the second hypothesis above. That is to say, syntactic recursion is more likely to be found in the grammars of languages spoken in societies associated with greater complexity.

One centrally important aspect of societal complexity mentioned earlier is modality. As already noted, syntactic recursion is more commonly found in writing than in speech; moreover, this tendency may be manifest in two distinct ways, online and conventionalized. To begin with, multiple embedding and associated syntactic recursion are more likely to occur in writing than in speech within a single language or even speaker. But in addition, within the same modality, languages with an overall higher rate of literacy are more likely to make use of multiple embedding and allow syntactic recursion. Evidence for this latter correlation was provided earlier by the contrasting results of the Google searches in English, with a higher rate of literacy, and Malay/Indonesian, which, as suggested in Gil (2009: 30), is characterized by a lower rate of literacy and correspondingly lower functional range of written communication.

Nevertheless, literacy is just one of an array of features generally associated with societal complexity, and there is good reason to believe that several of these other features may also be associated with a greater propensity for recursion. A

series of recent studies have demonstrated positive correlations between various aspects of societal and linguistic complexity. Thus, recent experiments by Raviv et al. (2019, 2020) and Raviv (2020) show that in artificial languages, larger speech communities create more highly compositional languages, entailing increased complexity in the domain of combinatorial syntax. Similarly, in sign languages, Meir et al. (2012) and Ergin et al. (2020) argue that an increase in the size of the signing community entails a greater degree of conventionalization. In phonology, Hay & Bauer (2007), Atkinson (2011), Wichmann et al. (2011) and Nettle (2012) argue that larger languages tend to have larger phonemic inventories than smaller languages. In the domain of metaphor comprehension, Gil & Shen (2021) present evidence to the effect that more highly complex polities tend to be associated with languages whose metaphors are of more complex directional structure. With respect to Tense-Aspect-Mood marking, Gil (2021) demonstrates that languages belonging to larger families, the product of demographic spread, are associated with more complex systems characterized by obligatory as opposed to optional marking. Finally, in the realm of basic clause structure, work in progress, some initial results of which are summarized in Gil & Shen (2019), shows that more highly complex polities tend to be associated with languages endowed with a greater degree of grammaticalization of thematic-role assignment. Admittedly, though, a range of other studies support an opposite negative correlation between societal and linguistic complexity. As argued by McWhorter (2018, 2005, 2011), Dahl (2004), Wray & Grace (2007), Lupyan & Dale (2010), Trudgill (2011) and others, larger political entities, typically associated with various modes of exoteric communication, and in particular imperfect adult second-language acquisition, are conducive to linguistic simplification, whereas smaller societies, generally characterized by more esoteric forms of communication, are fertile grounds for the accretion of linguistic complexity. A way of reconciling these apparently divergent results is proposed in Chen et al. (2024), who demonstrate, in a quantitative study based on data from the World Atlas of Language Structures, or WALS (Haspelmath et al. 2005), that many of the linguistic features associated with a positive correlation between linguistic and societal complexity tend to be of a syntactic rather than morphological nature; in addition, they speculate that such features will tend to rely more on procedural as opposed to declarative memory. Since recursion was not included in the features examined in WALS, it is not mentioned in the Chen et al. (2024) study. However, the generalizations emerging from that study strongly suggest that, as manifestations par excellence of syntactic complexity, syntactic subordination and recursion should also exhibit a positive correlation with societal complexity.

Viewed in the light of the covariance of societal and syntactic complexity, the facts regarding recursion in Malay/Indonesian, Pirahã and English, as summarized in Figure 5, make perfect sense. In terms of societal complexity, Malay/Indonesian occupies an intermediate position between Pirahã and English. Whereas Pirahã is spoken by a single small group of people with relatively little contact with the outside world, Malay/Indonesian, in all of its varieties, is spoken by well over two hundred million people and is an official language of four different countries. No wonder, then, that, unlike Pirahã, Malay/Indonesian allows for syntactic recursion, a fact that is true also for colloquial varieties such as Riau Indonesian. However, in contrast to Malay/Indonesian, English is a world language, a vehicle not only of national but also international communication; moreover, unlike Malay/Indonesian, its standardized versions are spoken natively by large populations of speakers. It is this difference that accounts for the massively greater frequency of subordination, a manifestation of syntactic recursion, in actual English usage, as contrasted with Malay/Indonesian.

Abbreviations

ASSOC	associative
HYP	hypocoristic
PRTC	participant

Acknowledgements

It is a pleasure to dedicate this paper to Dan Everett, in some ways a kindred soul. I am also grateful to Ted Gibson for inviting me to be part of this festschrift. For discussions of some of the issues dealt with in this paper I am indebted to Jim Hurford, Ray Jackendoff and Eva Wittenberg. Thanks are due also to Caleb Everett for his comments on an earlier draft. Some of the material in this paper was presented at the International Symposium on Malay/Indonesian Linguistics (ISMIL), in the 14th meeting in Minneapolis on 30 April 2010, the 19th meeting in Jambi on 13 June 2015, and at the 26th meeting in Honolulu on 19 May 2023; I am grateful to participants at those events for helpful comments and criticisms. Finally, the work reported on here would not have been possible without the generous support of the Department of Linguistics of the Max Planck Institute of Evolutionary Anthropology and its director Bernard Comrie.

References

- Aarne, Antti & Stith Thompson. 1961. *The types of the folktale: A classification and bibliography*. Helsinki: Suomalainen Tiedeakatemia.
- Atkinson, Quentin D. 2011. Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science* 332. 346–349. DOI: 10.1038/nature09923.
- Behaghel, Otto. 1932. *Deutsche Syntax: Eine geschichtliche Darstellung*, vol. IV: Wortstellung. Heidelberg, Germany: Carl Winter.
- Chen, Sihan, David Gil, Sergey Gaponov, Jana Reifegerste, Tessa Yuditha, Tatiana Tatarinova, Ljiljana Progovac & Antonio Benitez-Burraco. 2024. Linguistic and memory correlates of societal variation: A quantitative analysis. *PLOS One* 19(4): e0300838. DOI: <https://doi.org/10.1371/journal.pone.0300838>.
- Collinder, Björn. 1960. *Comparative grammar of the Uralic languages*. Stockholm: Almqvist & Wiksell.
- Cysouw, Michael. 2023. Hierarchical structure, A new terminological conundrum for martin. Paper presented at the Colloquium on Martin Haspelmath's 60th Birthday, University of Potsdam.
- Dahl, Östen. 2015. How WEIRD are WALS languages. Paper presented at the conference "Diversity linguistics: Retrospect and Prospect", Max Planck Institute for Evolutionary Anthropology, Leipzig. <https://zenodo.org/record/7473063> (16 March, 2024).
- Dahl, Östen. 2004. *The growth and maintenance of linguistic complexity*. Amsterdam: John Benjamins.
- Derbyshire, Desmond C. 1979. *Hixkaryana* (Lingua Descriptive Series 1). Amsterdam: North Holland.
- Deutscher, Guy. 2000. *Syntactic change in Akkadian: The evolution of sentential complementation*. Oxford: Oxford University Press.
- Dixon, Robert M. W. 1972. *The Dyirbal language of North Queensland*. Cambridge: Cambridge University Press.
- Dixon, Robert M. W. 1981. Wargamay, the Mpakwithi dialect of Anguthimri; Watjarri; Margany and Gunya; Tasmanian. In Robert M. W. Dixon & Barry J. Blake (eds.), *Handbook of Australian languages*, vol. 2, 1–144. Canberra: Australian National University Press.
- Eberhard, David M., Gary F. Simons & Fennig Charles D. (eds.). 2020. *Ethnologue: Languages of the world*. Twenty-third edition. Dallas, TX: SIL International. <http://www.ethnologue.com> (16 March, 2024).
- Englebretson, Robert. 2003. *Searching for structure: The problem of complementation in colloquial Indonesian conversation*. Amsterdam: John Benjamins.

- Ergin, Rabia, Limor Raviv, Ann Senghas, Carol Padden & Wendy Sandler. 2020. Community structure affects convergence on uniform word orders: Evidence from Emerging sign languages. In *The Evolution of Language: Proceedings of the 13th International Conference (Evolang13)*, 84–86.
- Everett, Daniel L. 1986. Pirahã. In Desmond C. Derbyshire & Geoffrey K. Pullum (eds.), *Handbook of Amazonian languages*, vol. 1, 200–325. Berlin: Mouton de Gruyter.
- Everett, Daniel L. 2005. Cultural constraints on grammar and cognition in Pirahã. *Current Anthropology* 46(4). 621–646.
- Everett, Daniel L. 2017. Grammar came later: Triality of patterning and the gradual evolution of language. *Journal of Neurolinguistics* 41(8). 133–165.
- Foley, William A. 2010. Clause linkage and nexus in Papuan languages. In Isabelle Bril (ed.), *Clause linking and clause hierarchy*, 27–50. Amsterdam: John Benjamins.
- Futrell, Richard, Laura Stearns, Daniel L. Everett, Steven T. Piantadosi & Edward Gibson. 2016. A corpus investigation of syntactic embedding in Pirahã. *PLOS ONE* 11(3). e0145289. DOI: 10.1371/journal.pone.0145289.
- Gil, David. 2000. Syntactic categories, cross-linguistic variation and universal grammar. In Petra Vogel (ed.), *Approaches to the typology of word classes, empirical approaches to language typology*, 173–216. Berlin: Mouton.
- Gil, David. 2005. Word order without syntactic categories: How Riau Indonesian does it. In Andrew Carnie, Heidi Harley & Sheila Dooley (eds.), *Verb first: On the syntax of verb-initial languages*, 243–263. Amsterdam: John Benjamins Amsterdam.
- Gil, David. 2006. Intonation and thematic roles in Riau Indonesian. In Chungmin Lee, Matthew Gordon & Daniel Büring (eds.), *Topic and focus, cross-linguistic perspectives on meaning and intonation*, 41–68. Dordrecht: Springer.
- Gil, David. 2009. How much grammar does it take to sail a boat? In Geoffrey Sampson, David Gil & Peter Trudgill (eds.), *Language complexity as an evolving variable*, vol. i, 19–33. Oxford: Oxford University Press.
- Gil, David. 2012. Where does predication come from? *Canadian Journal of Linguistics* 57. 303–333.
- Gil, David. 2013. Riau Indonesian: A language without nouns and verbs. In Jan Rijkhoff & Eva van Lier (eds.), *Flexible word classes*, 89–130. Oxford: Oxford University Press.
- Gil, David. 2017. Isolating-monocategorial-associational language. In Henri Cohen & Claire Lefebvre (eds.), *Handbook of categorisation in cognitive science*, Second Edition, 471–510. Oxford: Elsevier.

- Gil, David. 2020. What does it mean to be an isolating language? The case of Riau Indonesian. In David Gil & Antoinette Schapper (eds.), *Austronesian undressed: How and why languages become isolating*, 9–96. Amsterdam: John Benjamins.
- Gil, David. 2021. Tense–aspect–mood marking, language-family size and the evolution of predication. *Philosophical Transactions of the Royal Society B*. 376(1824). 1–16.
- Gil, David. 2023. Recent contact features in Northwest Papuan Malay. *NUSA: Linguistic Studies of Languages in and around Indonesia* 75.
- Gil, David & Yeshayahu Shen. 2019. How grammar introduces asymmetry into cognitive structures: Compositional semantics, metaphors and schematological hybrids. *Frontiers in Psychology - Language Sciences* 10. 1–17. DOI: 10.3389/fpsyg.2019.02275.
- Gil, David & Yeshayahu Shen. 2021. Metaphors: The evolutionary journey from bidirectionality to unidirectionality. *Philosophical Transactions B*. 376. 1–7.
- Givón, Talmy. 1979. *On understanding grammar*. New York, NY: Academic Press. 1–321.
- Grace, George W. 1987. *The linguistic construction of reality*. London: Croom Helm.
- Hale, Kenneth. 1979. On the position of Walbiri in a typology of the base. Ms. MIT.
- Hale, Kenneth. 1983. Warlpiri and the grammar of non-configurational languages. *Natural Language and Linguistic Theory* 1(1). 5–47.
- Haspelmath, Martin, Matthew Dryer, David Gil & Bernard Comrie (eds.). 2005. *The world atlas of language structures*. Oxford: Oxford University Press.
- Hay, Jennifer & Laurie Bauer. 2007. Phoneme inventory size and population size. *Language* 83. 388–400.
- Henrich, Joseph. 2020. *The WEIRDest people in the world, how the West became psychologically peculiar and particularly prosperous*. New York, NY: Farrar, Straus & Giroux.
- Henrich, Joseph, Steven J. Heine & Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and Brain Sciences* 33(2-3). 61–83. <https://hdl.handle.net/10419/43616> (16 March, 2024).
- Hurford, James R. 2011. *The origins of grammar, language in the light of evolution II*. Oxford: Oxford University Press.
- Jackendoff, Ray & Eva Wittenberg. 2014. What you can say without syntax: A hierarchy of grammatical complexity. In F. Newmeyer & L. Preston (eds.), *Measuring linguistic complexity*, 65–82. Oxford: Oxford University Press.

- Jackendoff, Ray & Eva Wittenberg. 2017. Linear grammar as a possible stepping-stone in the evolution of language. *Psychonomic Bulletin and Review* 24(1). 219–224.
- Karlsson, Fred. 2007a. Constraints on multiple center-embedding of clauses. *Journal of Linguistics* 43(2). 365–392.
- Karlsson, Fred. 2007b. Constraints on multiple initial embedding of clauses. *International Journal of Corpus Linguistics* 12.1. 107–118.
- Karlsson, Fred. 2009a. Origin and maintenance of clausal embedding complexity. In Geoffrey Sampson, David Gil & Peter Trudgill (eds.), *Language complexity as an evolving variable*, vol. i, 192–202. Oxford: Oxford University Press.
- Karlsson, Fred. 2009b. Syntactic recursion and iteration. In Harry van der Hulst (ed.), *Recursion and human language*, 43–67. Berlin: Mouton de Gruyter.
- Kirby, Kathryn R., Russell D. Gray, Simon J. Greenhill, Fiona M. Jordan, Stephanie Gomes-Ng, Hans-Jörg Bibiko, Damián E. Blasi, Carlos A. Botero, Claire Bowern, Carol R. Ember, Dan Leehr, Bobbi S. Low, Joe McCarter, William Divale & Michael C. Gavin. 2016. D-PLACE: A global database of cultural, linguistic and environmental diversity. *PLoS ONE* 11(7). e0158391. DOI: 10.1371/journal.pone.0158391.
- Koehn, Ed & Sally Koehn. 1986. Apalai. In Desmond C. Derbyshire & Geoffrey K. Pullum (eds.), *Handbook of Amazonian languages*, vol. 1, 33–127. Berlin: Mouton de Gruyter.
- Lupyan, Gary & Rick Dale. 2010. Language structure is partly determined by social structure. *PLOS ONE* 5(1). e8559. DOI: 10.1371/journal.pone.0008559.
- Mammeri, Mouloud. 1985. *L'ahellil de gourara*. Montreal, Canada: Éditions de la Maison des sciences de l'homme.
- McWhorter, John H. 2005. *Defining Creole*. Oxford: Oxford University Press.
- McWhorter, John H. 2011. *Linguistic simplicity and complexity: Why do languages undress?* Berlin: De Gruyter Mouton.
- McWhorter, John H. 2018. *The Creole debate*. Cambridge: Cambridge University Press.
- Meier, Ernst. 1851. *Deutsche kindereime und kinderspiele aus Schwaben*. Tübingen: Ludwig Friedrich Fues.
- Meir, Irit, Assaf Israel, Wendy Sandler, Carol Padden & Mark Aronoff. 2012. The influence of community on language structure: Evidence from two young sign languages. *Linguistic Variation* 12(2). 247–291.
- Nettle, Daniel. 2012. Social scale and structural complexity in human languages. *Philosophical Transactions of the Royal Society B: Biological Sciences* 367(1597). 1829–1836.

- Newell, Wiliam Wells. 1905. The passover song of the kid and an equivalent from new England. *The Journal of American Folklore* 18(68). 33–48.
- Pawley, Andrew. 1993. A language which defies description by ordinary means. In William A. Folley (ed.), *The role of theory in language description*, vol. 69, 87–129. Mouton de Gruyter Berlin.
- Pullum, Geoffrey K. 2024. Daniel Everett on Pirahã syntax. In Edward Gibson & Moshe Poliak (eds.), *From fieldwork to linguistic theory: A tribute to Dan Everett*, 23–74. Berlin: Language Science Press. DOI: 10.5281/zenodo.12665907.
- Pullum, Geoffrey K. & Barbara C. Scholz. 2010. Recursion and the infinitude claim. In Harry van der Hulst (ed.), *Recursion in human language*, 113–138. Berlin: Mouton de Gruyter.
- Raviv, Limor. 2020. *Language and society: How social pressures shape grammatical structure*. Radboud University, Nijmegen. (Doctoral dissertation).
- Raviv, Limor, Antje Meyer & Shiri Lev-Ari. 2019. Larger communities create more systematic languages. *Proceedings of the Royal Society B* 286(1907). 1–9.
- Raviv, Limor, Antje Meyer & Shiri Lev-Ari. 2020. The role of social network structure in the emergence of linguistic structure. *Cognitive Science* 44. e12876. DOI: 10.1111/cogs.12876.
- Ross, Malcolm D. 2001. Contact-induced change in Oceanic languages in North-West Melanesia. In A. Y. Aikhenvald & Robert M. W. Dixon (eds.), *Areal diffusion and genetic inheritance*, 134–166. Oxford: Oxford University Press.
- Sandalo, Filomena, Cilene Rodrigues, Tom Roeper, Luiz Amaral, Marcus Maia & Glauber Romling da Silva. 2018. Self-embedded recursive postpositional phrases in Pirahã: A pilot study. In Luiz Amaral, Marcus Maia, Andrew Nevins & Tom Roeper (eds.), *Recursion across domains*, 279–295. Cambridge: Cambridge University Press.
- Slobin, Dan I. 1996. From ‘thought and language’ to ‘thinking for speaking’. In John Gumperz & Stephen Levinson (eds.), *Rethinking linguistic relativity*, 70–96. Cambridge: Cambridge University Press.
- Smailus, Ortwin. 1975. *Textos mayas de belice y quintana roo*. Berlin: Gebr. Mann Verlag.
- Thompson, Stith. 1946. *The folktale*. Berkeley, CA: University of California Press.
- Thurston, William R. 1987. *Processes of change in the languages of North-Western new Britain* (Pacific Linguistics B. - 99). Canberra, CA: The Australian National University.
- Trudgill, Peter. 2011. *Sociolinguistic typology: Social determinants of linguistic complexity*. Oxford: Oxford University Press, Oxford.

- Voorhoeve, C. L. 2010. A remarkable chain tale from new Guinea. In Kenneth McElhanon & Ger Reesink (eds.), *A mosaic of languages and cultures: Studies celebrating the career of Karl J. Franklin*, 408–415. Dallas, TX: SIL International.
- Wichmann, Søren, Taraka Rama & Eric W. Holman. 2011. Phonological diversity, word length and population sizes across languages: The ASJP evidence. *Linguistic Typology* 15. 177–197.
- Wray, Allison & George W. Grace. 2007. The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form. *Lingua* 117. 543–578.