Introduction: A Map of the Theoretical and Empirical Issues

Luiz Amaral, Marcus Maia, Andrew Nevins, and Tom Roeper

1 Recursion: Easy to Describe, Not Always as Easy to Find

Recursion is defined in a number of ways, from what Pinker and Jackendoff (2005:203) call "a procedure that calls itself, or … a constituent that contains a constituent of the same kind," or in the words of Rodgers and Black (2004), "a data structure that is partially composed of smaller or simpler instances of the same data structure."[1] Classic examples of recursion in syntactic theory arise with rewrite rules of the form developed in Post (1943) and subsequently familiar throughout all syntactic theory that include, among many other rules of course, the following key interacting rules:

(1) a. S → NP VP
b. NP → that S
c. VP → V NP
d. NP → the N

The combination of these rules yields indirect recursion, as expansion of (1a) into the VP rewrite rule in (1b) yields an infinite set of sentences of the form *The dog thought that the cat saw the rat* or *The dog thought that the bird said that the cat saw the rat*. With only the four rules in (1), an infinitude of sentences can be derived, and as Bar-Hillel (1953:164) pointed out, in the recursive analysis of such sentences with expansions like (1b), "we had to move from 'sentence' to 'nominal', then back to 'sentence', and finally once more to 'nominal'."

Similar properties arise once conditionals are added, as in (2), which draws on examples formulated in Chomsky (1957):

(2) S → if S then S

[1] The editors thank John Goldsmith, Andrea Moro, Ian Roberts, Pieter Seuren, and Charles Yang for very helpful comments and suggestions, many of which have made their way into this introduction.

Thus we derive *If the cat saw the rat then the dog told the bird* and even self-embedded conditionals like *If if the cat saw the rat…* Indeed, one can apply such a formalism to NPs themselves:

(3) a. NP → the N'
b. N' → N PP
c. N' → N
d. PP → P NP

This set of rewrite rules will yield *The book on the table near the mat* and so forth.

Taking what these have in common, we may say that an algorithmic generative rule system is recursive if the output of a given rule R2 (like 1b or 3d) contains a symbol or sequence of symbols Z that is also part of the input of a rule R1 (like 1a or 3a), such that R1 immediately or remotely generates Z. Note that R1 and R2 may be identical, as in (2).

This is an incredibly powerful property for languages to have. Husserl's (1913) *Logical Investigations* wrote of a way of obtaining "a boundless multitude of further forms" (as translated in Goldsmith and Laks 2016) by the recursive combination of propositions in this way, a characteristic of human language whose nature has been studied in virtually all subsequent work on the topic. Notably, however, the discussion of recursion in phrase structure in the first few decades of research on the topic focused almost exclusively on recursion involving sentential embedding (see Graffi 2015 for discussion). However, recursion, whether of the center-recursion kind in (2), or the edge-recursion kind in (3), can be found across domains beyond sentential embedding alone, and once PP and NP recursion are included, it is potentially found in some form or another in all languages. This depends, of course, on the formalism used to express it grammatically, as a language that allows, say, up to four adjectives (*the shiny bright expensive green bottle*) has a more parsimonious grammatical description than one that needs to add a new phrase structure rule for every adjective that is added (as in (5)):

(4) a. NP → the N'
b. N' → Adj N'
c. N' → N

(5) a. NP → the N
b. NP → the Adj N
c. NP → the Adj Adj N
d. NP → the Adj Adj Adj N
e. NP → the Adj Adj Adj Adj N

Nonrecursive formalisms (such as (5)) *could* be written if, say, a corpus of a language never found more than four adjectives in a row, and a linguist who

decided to model a grammar based on this insisted such a corpus was truly representative of the language. Conversely, one might question whether, indeed, the limit to four adjectives was found specifically because this was a finite corpus, based on limited genres like monologue narratives with no real-time conversational interactions, instead of being from a wide range of dyads, facing goal-oriented demands, or contextually varied situations of wooing, scolding, or planning. If elicitation tasks reveal that narrative-based corpora present only a limited sample of the language (see Davis, Gillon, and Matthewson 2014), then formalisms such as (5) will be inferior to the coverage and compactness of those in (4).

How to tie the widespreadness (or lack thereof) of a given recursive pattern in actual usage to its formalism? It is well established (see Chomsky and Miller 1963; Gibson and Thomas 1999) that center-embedding becomes very difficult to process, to the point of rendering grammaticality judgments difficult after three embeddings. Clearly, therefore, recursion in this domain involves more processing costs. Focusing on this kind of recursion would make one doubt its existence. By contrast, focusing on possessor recursion might make one quite confident that recursion is alive, kicking, and easy to process; Lima and Kayabi (this volume) find, for example, that Kawaiwete-speaking children correctly answer questions like "What is Pedro's friend's brother's basket's color?" Comparison of recursion *across domains* is therefore crucial, and particularly across a range of languages and with a range of methods, especially when grappling with the question of whether recursion is at the center of every language (as many interpret Hauser, Chomsky, and Fitch 2002).

It becomes immediately clear that no matter how easy it may be to write down a recursive set of rules such as (1)–(3), the full-blown use of recursion does not occur as widely as we might expect, as reflected in typological skewedness, language-specific limitations, real-time processing costs as measured behaviorally, with eye-tracking or brain-imaging, and finally in its acquisition profile. Children's acquisition of clausal embedding of the type generated in (1), studied by Bloom et al. (1989), for example, has been argued to be somewhat limited in its usage at an early age. Diessel (2004), in a thorough study of the uses of clausal embedding introduced by "I think that…" in child language, argues that in such cases, the "apparent matrix clauses are nonassertive: rather than denoting a mental state or an act of perception, they function as epistemic markers, attention getters, or markers of illocutionary force" (p.3). In fact, children sometimes omit the complementizer 'that' in even German and French, languages in which the complementizer is obligatory for adults, and frequently use "I think" as a sentence-final parenthetical. In such a characterization, the relationship between expressions such as "I think" and the proposition they introduce does not follow the hallmarks of true embedding, where embedded propositions are syntactically and semantically integrated in the matrix clause and marked as dependent structures that are

formally incomplete without the matrix clause. Instead, one might treat these utterances with a rewrite rule:

(6) S → NP think that NP VP

A rule such as (6) is nonrecursive in the sense that it explicitly rules out double-embedding. Diessel's argument is that younger children start out with rule systems such as (6), only gradually later giving way to a revision of the type that yields (1).

Along somewhat similar lines, not all *languages* employ direct clausal embedding of the type that (1) would derive. Levinson (2013), for example, claims that recursion in Kayardild is limited to one level deep, precisely as (6) would restrict it, although for reasons of case morphology related to being polysynthetic. Indeed, it is sometimes claimed that the type of grammatical reanalysis that is responsible for children's transitions from paratactic embedding structures as in (6) to fully fledged embedding structures that would be generated by (1) finds parallels at longer timescales, in the kind of grammaticalization that diachronically yields clausal embedding from formerly paratactic structures. Givón (2009) shows how this kind of reanalysis takes place with examples from Bambara, Hittite, Germanic, and Uto-Aztecan, arguing that one of the parallel components in this reanalysis is when the previously adjoined material now "falls under a single merged intonation contour with the main clause" (p.202). Similarly, Hale (1976:78) argued that:

> [in] a large number of Australian languages, the principal responsibility for productive recursion in syntax is shouldered by [an] adjoined relative clause. It is typically marked as subordinate in some way, but its surface position with respect to the main clause is marginal rather than embedded.

Similarly, Nordlinger (2006) argues that structures such as *they drink grog, they'll fight* in the Australian language Wambaya are fully ambiguous between coordinated and subordinated (if-then) relations (although she contends that the subordinate construal may be forced by prosody). Nordlinger's discussion, alongside a detailed overview of Australian language data by Legate, Pesetsky, and Yang (2014), makes clear that it is far from correct to say that subordination is lacking in these languages, but nonetheless that this kind of recursion is not as freely used as it is in English, or in comparison to adjectival recursion in the same languages. If anything, then, while it is definitely too radical and simplistic to say that any language 'lacks' recursion, its distribution parallels that of, say, interdental fricatives like /θ, ð/ in English: positionally limited (found only before *r* in clusters), not often found with more than one instance (*thither*), and notoriously difficult in L1 and L2 acquisition.

What leads to these markedness-like restrictions on 'a constituent that contains a constituent of the same kind,' and how do they line up with different

kinds of recursion? Why is it that some languages are more restrictive with CP recursion of the type exhibited in (1) than they are with PP recursion exhibited in (3)? Why do some languages allow recursive noun-noun compounding (e.g., *comic book club*) while others do not (Roeper and Snyder 2004)? Recursion looks like a candidate for the application of Jakobson's (1941) criteria of markedness as compellingly linked in his book *Kindersprache, Aphasie, und Allgemeine Lautgesetze*, in which he argued that marked phonological structures are rarer cross-linguistically as well as later to develop in child language. Again, these questions must be asked with respect to a much richer range of structures than center-embedding relative-clauses or sentential embedding of speech reports. What is the typological, acquisitional, and processing profile of DP recursion? Of PP recursion? Of causatives, evidentials, imperatives, and other underexplored XPs potentially allowing being pushed inside 'a constituent of the same kind'? Coming back to (3), with the case of PPs for example, interpretational issues arise. How do we know whether the interpretation of *the book on the table next to the mat* is really one of *next to the mat* modifying *the book*, rather than *the table*? How do we know that, for example, real-time instances of these are implemented in comprehension by recursion, rather than iteration? How does cross-linguistic transfer of recursive structures take place in L2 acquisition (see Nelson 2016 for PP recursion across Spanish-English L2)?

Moreover, certain kinds of XPs also naturally lend themselves to embedding more easily than others. For example, imperatives like *Go fetch the water* are crosslinguistically less likely to be embeddable. On the other hand, the kind of recursion found with possessors, as in *My sister's husband's boss got us the tickets*, are extremely natural in adult English. (And as mentioned, in Lima and Kayabi's chapter in this volume, up to four levels of possessor recursion are easily processed by Kawaiwete-speaking children.) In other languages, however, restrictions are imposed on such structures. As argued in Nevins, Pesetsky, and Rodrigues (2009b), such restrictions may be in some instances morphosyntactic: languages with certain kinds of morphological systems for case or finiteness may disallow one instance of finite morphology to be c-commanded by another. Rizzi (2013) contends that while the formal operations yielding recursion are fully available as part of general computation, its applications are modulated by the specific properties of the lexical items it acts upon. Just such an investigation of which properties, and how they modulate the possibilities of recursion, is precisely at stake in this current volume.

How does the possibility of recursive operations within particular domains come to change over various timescales – the timescale of a child that integrates language use with other cognitive systems, for example? Consider sentences such as (7):

(7) What did the girl say she bought?

De Villiers (1999) showed that when children under four years are asked to answer questions like this one, they tend to answer with what the girl actually bought rather than what she *said* she bought, which of course might be different. According to de Villiers, these children have not yet analyzed (4) in terms of an embedding syntax and semantics; indeed, she argues that cognitive development and language acquisition are mutually dependent, and that the development of faculties such as Theory of Mind and false belief go hand in hand with the analysis of (7) as fully fledged embedding. The dramatic connections between lack of subordination and false belief – the question of when children realize the "opacity" of complement sentences – parallels the evolution of subordination types along the acquisition path. These, in turn, mirror the variation across languages discussed above. A subpath of stages, with stepwise semantic changes as well (see Roeper and de Villiers 2011; de Villiers et al. 2012), precedes the full form of recursive complementation that syntactically and semantically represents false belief. These types of subordination are stepping stones to the full subcategorization of clauses that enables recursive structure for multiple points of view. In their contribution to this volume, Roeper and Oseki argue that there is an acquisition path from coordination to indirect recursion.

While recursion across the domains of sentential embedding, prepositional phrases, causative structures, possessors, and relative clauses can all be described formally with the same means (e.g., the rewrite rules in (1)–(3)), their distribution across and within the world's languages is clearly not equal. With a focus on cross-linguistic diversity, this volume includes experiments on PP recursion, possessive recursion, relative clause recursion, adjective recursion, sentential recursion of both tensed and non-finite clauses, and discourse recursion. This will allow us to begin to consider new kinds of questions: are there clusters of recursive structures that are reflected in typology, dialects, or language change? Are they acquired in a systematic way with one kind triggering another (as discussed in Roeper and Oseki's chapter)? Finally, what interfaces do they have with morphology, parametric variation, and lexical representation?

Our goal in this volume is to bring new data and emerging research methodologies from a gamut of less familiar languages to this study. This book aims to address a host of topics about recursion woven together across different dimensions of linguistic research: formal analysis, theoretical exploration, experimental fieldwork, and several methodologies (intuition, comprehension experiments, event-related potentials, and reaction time studies). Recursion is held up against its interaction with reference, evidentiality, second-order beliefs, and prosody across domains and latitudes, and it is compared to non-recursive structures (e.g., the 'coordination default') across fifteen languages,

thereby informing the question of its distribution, processing, and restrictions across grammatical domains and across grammars.

2 Recursion across Latitudes

This book presents analyses of recursive constructions in a broad array of languages representing a great areal, typological, and genetic diversity, spanning wide latitudes. Fifteen languages are examined in the eighteen chapters of this book. Dutch, English, Portuguese, and Japanese are studied, alongside eleven Amerindian languages of South America, coming from six distinct language families or genetic affiliations.

The region of the world that is currently most prominent in contemporary discussions of recursion is lowlands South America (east of the Andes), in particular the regions surrounding the Brazilian Amazon, in large part due to the claims that arose in Everett (2005) that Pirahã, an indigenous language of Brazil, has 'no embedding,' and the subsequent debates that arose about (a) what, in fact, are the right kind of empirical tests to be conducted to make such claims (see Sauerland and Sandalo et al. in this volume for new kinds of experimental inquiry for this language) and (b) whether, if indeed, Pirahã were to be missing recursion in one particular domain (e.g., CP), what consequences this might have for developing a theory of markedness of recursive structures. Given particularly the linguistic diversity of Brazil, the unfamiliarity of many of these languages to mainstream debates on recursion, and the vibrant presence of a range of interdisciplinary methods applied by researchers working on them, their contribution to the questions laid out above is extremely valuable.

We turn, therefore, from a mapping out of linguistic domains to a necessary introduction to the distribution of recursion across latitudes as reflected in the research reported on languages within Brazil.

Brazil has, at present count, around 150–160 indigenous languages (Moore 2011). The present volume considers phenomena from ten living Brazilian indigenous languages (depicted in Figure 0.1), as well as Tupinambá, an extinct language spoken on the coast of Brazil when the first colonizers arrived. These comprise thirteen chapters in the book. Alongside these are English, Portuguese, Japanese, and Dutch, which are examined in the other five chapters, constituting the total eighteen chapters of the book. In addition to their wide areal distribution, ranging from Kotiria and Wapichana in the northwest and north regions of the country, Pirahã in the south Amazon area and Karitiana in the western state of Rondonia, to Guarani in the south and southeast, with a concentration of languages in Central Brazil (Karajá, Kawaiwete, Kĩsêdjê, Kuikuro), the languages studied in the book are also representative of

Figure 0.1 Living languages with approximate locations

some of the main linguistic stocks and families in Brazil: Tupian, Macro-Jê, Carib, Arawak, and Tukano (alongside an isolate, Pirahã).

The Tupian stock, the largest South American genetic group, is represented by languages from both its western branch (Karitiana, the last surviving

language of the Arikem family) and eastern branch (Kawaiwete, Tenetehára, Mbyá Guarani), all belonging to the Tupi-Guarani family, the largest family in the stock. Damaso Vieira analyzes recursive constructions in Tupinambá, an extinct language of the Tupi-Guarani family, drawing on data registered in the sixteenth and seventeenth centuries.

Languages of the second largest genetic group in Brazil, the Macro-Jê stock, are also well represented in the book: Kĩsêdjê, a member of the Jê family, the largest family in the stock, and Karajá (Karajá family). The Carib family, which contains the second largest number of languages in a single family, is represented by Kuikuro, a language classified in the Xinguan southern branch of Carib. The Arawak family is represented in the book by Wapichana, the North Arawakan language with the largest number of speakers. Kotiria represents the Eastern Tukano family, spoken in the northwestern Amazon basin. Finally, Pirahã, which is examined in three of the chapters in the book, is the last surviving language of the small Mura family, spoken in the south Amazon area.

Even though the Brazilian indigenous languages investigated in the book present different levels of vitality according to UNESCO criteria, there is a clear consensus among linguists working on indigenous languages in Brazil that all of these languages are classified at least as "vulnerable" by UNESCO's endangerment criteria. Spoken by a mostly monolingual population, Pirahã has a high level of intergenerational transmission, but the sheer low number of speakers (less than 400) makes it vulnerable. With a population of around 7,000 speakers in Brazil, Wapichana, on the other hand, is considered "definitely endangered" by UNESCO, based on the fact that less than half of the population can speak the language (see Moore 2011) and that at least 80 percent of the population is bilingual in Portuguese. While most of the languages discussed are largely understudied, the research presented in this book, with the exception of the study on extinct Tupinambá, are based on first-hand data, collected by the authors of the chapters themselves.

All of the studies presented in the book were developed based on a cross-comparative methodology, in which theoretical questions guide data collection. Some of the studies engage in the precision offered by experimental methods, a new endeavor which is being called "experimental fieldwork," facing the challenge of bringing together crucial dimensions of linguistics such as theory of grammar, psycholinguistic methods, and fieldwork procedures in order to attempt to uncover grammatical processes that could never be discovered solely on the basis of corpus building. Thus, while our focus in this section has been on the typological diversity of the languages covered here, of equal importance in our organization of this volume is to convey the range of similar methodologies that can and should be applied across them. There are thus direct connections between the issues raised and Theory of Mind, PP recursion, and coordination and subordination in parallel investigations with English,

Portuguese, Japanese, and Dutch found throughout this book (see especially the contributions by Terunuma and Nakato, Hollebrandse, and Pérez-Leroux et al.). At the same time, the Brazilian languages represented herein involve empirically foci that are largely absent from well-studied European languages, e.g., evidential marking, switch-reference marking, and embedded imperatives.

3 Recursion and Embedding across Domains

This volume is organized into four grammatical domains in which recursion is examined. Clearly, at some points there may be intersections or transversal connections possible across individual chapters within distinct sections.

3.1 Speech Reports, Theory of Mind, and Evidentials

The formal property of sentential embedding is invariably linked to the semantic and illocutionary types of elements they link up. Most canonically, these are speech reports, which involve reporting the beliefs or speech acts of others, and thus intersect with questions of Theory of Mind, second-order beliefs, the cognitive development of Theory of Mind in children, and evidential reporting more generally.

Sauerland's chapter examines speech reports such as *Toi said that he has been to the moon*, which can be used to distinguish between subordination and coordination structures for speech reports. Specifically, if the relationship between the embedded proposition and the higher attitude verb is one of subordination, then the sentence as a whole can be judged as true. However, if Pirahã really lacked embedding, then sentences of this sort would actually be coordination structures, akin to *Toi spoke, and he has been to the moon*. The chapter discusses the result of an experiment conducted with sixteen Pirahã speakers, who as a whole end up distinguishing subordination (*Toi said that he has been to the moon*) from coordination (*Toi spoke, and he has been to the moon*), where the former as a whole can be judged as true at the same time that the latter is judged as false. The results therefore provide a new empirical base for the conclusion that Pirahã grammar contains at least one level of embedding, and moreover outline a technique for the study of speech reports that can be straightforwardly employed and replicated in experimental fieldwork situations with relatively understudied languages.

A further challenge, only recently beginning to be studied experimentally, is the extent to which one can trace the distribution and development of second-order belief ascription, as broached in the chapter by Hollebrandse, focusing on sentences like *The judge knows that the jury thinks that Malcolm is guilty*. These structures are particularly interesting for the subordination versus coordination dichotomy because, as Hollebrandse argues, there is virtually no way to