Language as an emergent system.

K. David Harrison and Eric Raimy
Swarthmore College

Human language is a biological system: All humans are neurologically predisposed to acquire whatever language they are exposed to in their early years. Language itself is socially transmitted. A rich interaction between genetic (brain) structures and social behavior gives rise to what linguists call a "grammar", which is the knowledge (part innate, part learned) of language complexity found inside the brain of every speaker. The task of Linguistics, properly a sub-field of Neurobiology in our opinion, is to describe and explain the complex patterns found in all languages. This apparent complexity, with its simpler underlying components, make language an emergent system *par excellence*. The paradigm of emergence set forth in this volume is very well-suited to the study of Linguistic patterns on at least two distinct (though interrelated) levels.

While we believe there is no exact diagnostic that can characterize an emergent system, there is a consensus that emergence manifests itself as complexity at a global level that is not locally specified or only very weakly so!. There is also a 'surprise' factor when complex behavior is not what we expect, given our (limited) knowledge of the underlying components.

Language learners, beginning from birth and on through the language learning years up to age seven or so, are confronted with massive amounts of complexity, but apparently are not surprised. In fact, they behave as if they expect an emergent system. That is, they never seem to assume that they must memorize verbatim all linguistic forms that they hear. Thus, they never represent in their brain the world exactly as it occurs. What they assume instead is that the complexity they are hearing must be produced by the interaction of simpler underlying mechanisms. Languages thus always end up being represented abstractly in the brain, we will claim.

Children acquiring language are 'emergonauts' in that they assume emergence everywhere, and will go for the abstract representation rather than the most direct, literal one. We attribute children's assumption of 'emergence' to the fact that the language faculty of the human brain contains statistical pattern detectors focused on analyzing language. In other words, the brain is built to detect and extract patterns. Although we make this assumption, the actual endeavor is an empirical research program which seeks to find answers about what types of statistics the language faculty is able to compute, and over what kinds of representations it computes them. Additionally, once we have useful answers to these particular questions we can then begin to address whether the statistical pattern detectors or the representations that the statistics are computed are unique to the language faculty, or more general-purpose. To demonstrate this view of the human language faculty we will discuss how emergent patterns in the pronunciation of different morphemes are part of the process of language acquisition in Tuvan, a language spoken in central Siberia. Our analysis extends to all human languages.

Emergent properties of morphology

Language learners are confronted with a dizzying array of variation in linguistic forms. Their task, made possible by innate statistical pattern detectors, is to map surface complexity of speech forms onto a more abstract set of representations to be memorized. Once they have the right representations in place, speakers have achieved 'competence', the ability to both parse and generate nearly limitless numbers of completely novel words, phrases and sentences.

As an example of the type of surface complexity facing learners, we present the case of suffix morphemes in Tuvan, a language of Siberia. A morpheme is the pairing of a sound and meaning into an atomic unit that is stored in long term memory, in the mental lexicon. Allomorphy refers to a chameleon-like quality some morphemes exhibit in taking on different phonological shapes to match different environments. In other words, a single memorized morpheme may assume one of multiple different related pronunciations when a user actually utters the morpheme. The task for the learner confronted with a suspected case of allomorphy is to decide whether there is just a single morpheme to be memorized and its chameleon like behavior can be accounted for by general rules of the language or if there are many distinct animals present which must be memorized separately without any generalizations to be had. The decision is aided by more general knowledge of Phonology, the speakers knowledge about what sound patterns are present in a language and how sounds affect other sounds when they appear in a particular environment. But the fact that this decision must be made at all by a learner cannot be derived simply from the complexity of the data she is exposed to or knowledge of sound patterns. Instead, we believe language learners face this question because the underlying pattern detection component of the language faculty forces them to do so. The system needs to derive abstractness from complexity so that language, with its infinite combinatorial possibilities, can be stored in finite brians.

We will begin our example on how a learner identifies allomorphy patterns from the surface forms of language by considering two suffixes in Tuvan, a Turkic language spoken by nomadic herders in Siberia. The first suffix we will consider is the plural marker, added to nouns, which has eight distinct allomorphs as presented in (1) (the suffix is shown in boldface).

(1) Tuvan plural suffix with eight allomorphs

Noun + plural suffix	meaning
*teve-**ler***	'camels'
*ulu-**lar***	'dragons'

*xep-**ter***	'clothes'
*at-**tar***	'names'
*xerel-**der***	'sunbeams'
*aal-**dar***	'campsites'
*xem-**ner***	'rivers'
*xam-**nar***	'shamans'

The second example we will consider is the sixteen allomorphs of the Tuvan adjectival suffix presented below in boldface font (2).

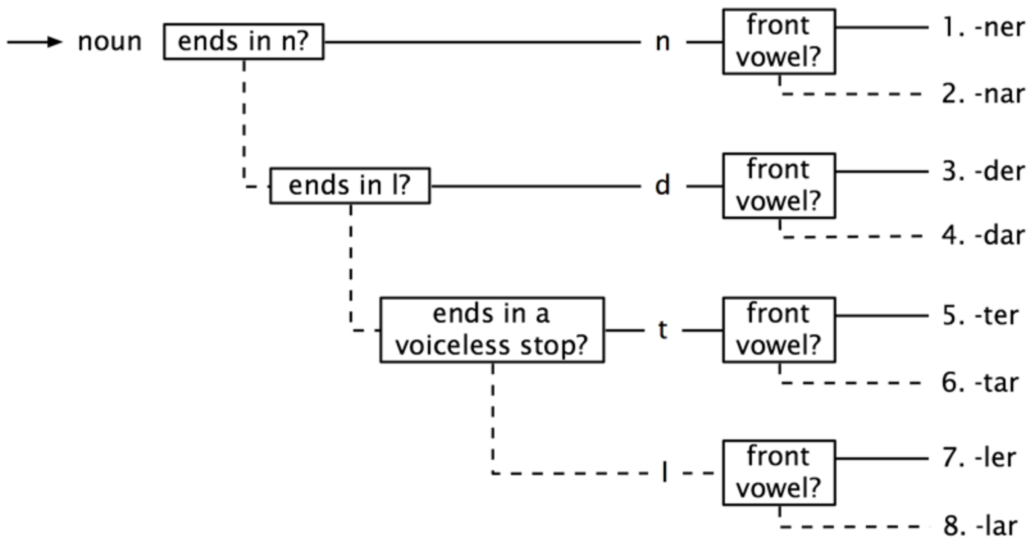
(2) Tuvan adjective suffix with 16 allomorphs

*teve-**lig***	'having a camel'
*böry-**lüg***	'having a wolf'
*ada-**lɨg***	'having a father'
*ulu-**lug***	'having a dragon'
*xep-**tig***	'having a clothing'
*üš-**tüg***	'having a three'
*àt-**tɨg***	'having a horse'
*quš-**tug***	'having a bird'
*xerel-**dig***	'having a beam of light'
*xöl-**düg***	'having a lake'
*aal-**dɨg***	'having a campsite'
*mool-**dug***	'having a Mongol'
*xem-**nig***	'having a river'
*xöm-**nüg***	'having leather'
*xam-**nɨg***	'having a shaman'
*qum-**nug***	'having some sand'

The question at hand is what do learners do when they are presented the sets of allomorphs in (1) and (2)? One important aspect of this data that must be commented on is that although there are many different allomorphs of the 'plural' and 'adjectival' suffix in Tuvan, learners do not have any problem in producing them and understanding that any and all of the allomorphs for 'plural' and 'adjective' have exactly the same meaning. Once we understand this point, we can sharpen our question to not what do the learners do with these allomorphs *per se*, but whether the learner makes any generalizations about the distribution of the allomorphs. More specifically, does the learner memorize eight distinct forms of the 'plural' in (1) with generalizations on what type of noun should occur with each allomorph? Or does the learner form some other type of generalization?

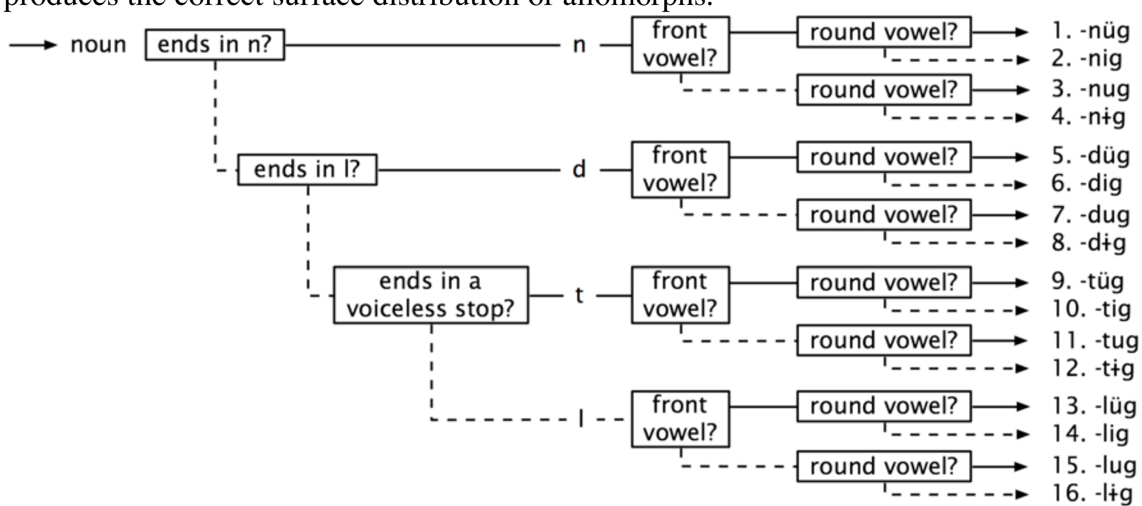
An important aspect of the distribution of the allomorphs in (1) and (2) is that the occurrence on different nouns is predictable based on what the phonological (sound) content of the noun is. We assume that this predictable aspect of the different allomorphs is statistically highly salient and thus readily identifiable by the learner. In a first pass of encoding, the predictable nature of the allomorphs for the 'plural' and 'adjective' suffixes in Tuvan can be represented by the decision trees in (3) and (4)

The decision tree for 'plural' allomorphs in (3) encodes a first draft of the generalizations about the distribution of these allomorphs in the following way. First, the learner has the assumption that there are eight distinct allomorphs present and what is occuring is the selection of which allomorph to use with what type of noun. The tree shows this decision process by beginning at the left edge, where the learner must answer a series of questions about the last sound of the noun. The first set of decisions focuses on the phonological content of the last sound of the noun. The first branch queries whether the noun ends in a 'nasal sound' (e.g. 'n', 'm' for the data set in (1) and (2)). If the answer is 'yes' then the learner has narrowed down the possible allomorphs to either #1 or #2 and the next query is about the last vowel in the noun. If this vowel is a 'front vowel' (i, e, y or ö for our data sets) then the learner selects allomorph #1 in (3) which is 'ner' and if it is not a 'front vowel' then allomorph #2, 'nar', is selected. The correct surface distribution of all eight plural allomorphs is produced by this decision tree following the rubric of a solid line indicates a 'yes' answer to the decision box and a dashed line indicates 'no'.



(3) Decision tree for ‘plural’ allomorphs

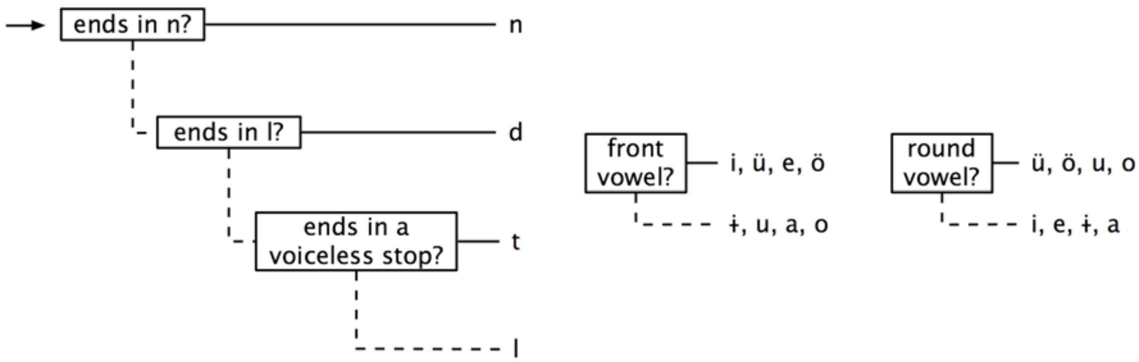
The decision tree in (4) for the ‘adjective’ allomorphs works in the same manner and produces the correct surface distribution of allomorphs.



(4) Decision tree for ‘adjective’ allomorphs

Although the decision trees in (3) and (4) represent enough information to predict the occurrence of the different allomorphs on different nouns, we do not believe the learner stops here. If we consider the content of the two decision trees in (3) and (4) an extremely important characteristic emerges. There are patterns within the decision trees themselves that learners detect. Specifically, we see these patterns as informational redundancies in the decision trees. One large redundancy that can be seen is that the ‘plural’ decision tree in (3) is wholly contained within the ‘adjectival’ decision tree in (4). The only difference between the two decision trees is the additional bifurcation based on whether the vowel in the noun is a ‘round’ (e.g. u, ü, o, ö) or ‘non-round’ (e.g. i, y, e, a) vowel. This bifurcation occurs in the ‘adjective’ tree but not the ‘plural’ tree and the learner takes this distribution as evidence to extract the 'roundness' sub-tree as a separate generalization.

The remaining decision tree which consists of the 'consonant' and 'front vowel' decisions can be further analyzed also when we consider the fact that Tuvan also contains suffixes which show the allomorphy in the vowel of a suffix but not the consonantal allomorphy. The possessive (3rd person) suffix in Tuvan has the following 4 allomorphs: /i/, /y/, /u/ and /ü/. For example, *nom* means 'book' and *nom-u* means 'his book'. Based on this information the learner can separate the 'consonant' decision from the 'front vowel' decision. This last extraction provides the three decision trees that will generate all of the thirty-two possible forms of the 'plural', 'adjective' and 'possessive' suffixes in Tuvan.



(5) Three decision trees for Tuvan: (a) 'consonant decision', (b) 'frontness decision', (c) 'roundness decision'.

At this point, there are no further patterns to be extracted from the decision trees. Accordingly, the generalizations are as 'simple' as possible and this result has been driven by the statistical analysis of patterns in the data. There are two interesting aspects of this current situation. The first is that the three decision trees coincide with a traditional linguistic analysis of Tuvan having three distinct phonological processes of 'consonant dissimilation', 'backness harmony' and 'roundness harmony' (Anderson and Harrison 1999). Each of these three processes are considered to be phonological (and not morphological) because the queries in each decision tree only refer to the sounds found in the noun that suffix is attaching to. The 'consonant dissimilation' decision tree affects the sound /l/ in Tuvan. The 'roundness harmony' decision tree affects the set of 'high' vowels in Tuvan and finally the 'backness decision' tree affects all vowels in Tuvan. At this point we can now understand how the learner identifies the allomorphs for the 'plural' and 'adjective' suffixes in (1) and (2) as chameleons. The patterns that can be extracted from the distribution from these allomorphs do not coincide with lexical or morphological information. Indeed the patterns range over the entire distribution of sounds in Tuvan. Because the pattern is not restricted to a particular word or morpheme, the learner need memorize just one of the surface allomorphs for the 'plural' and 'adjectival' suffixes. He may then allow the decision trees in (5) to modify the memorized morpheme to match specific sound environments.

A second interesting aspect of the decision trees in (5) is that relatively trivial transitional probability based statistics are the only thing needed to identify these generalizations. It is well documented that infants as young as 8 months are able to calculate transitional probabilities (Aslin, Saffran and Newport 1998). So we are confident that we are not making unreasonable claims about the statistical abilities of humans. One aspect of statistical learning that is often glossed over but we feel is a crucially important aspect of our analysis of Tuvan here is the question of what is done with the statistical knowledge gained by the learner. Statistical knowledge is useless in building words unless a decision is made based on it. We believe that the statistical analysis of the distribution of the 'plural' and 'adjective' suffixes in Tuvan provides the basis for the learner to make sensible decisions about what to memorize as the underlying, mental representations for the morphemes in question and what phonological processes are present in Tuvan. Although the statistical analysis provides the source of the knowledge about the distribution of allomorphs in Tuvan, the learner makes a decision about whether to memorize a static representation or not. It is the the static generalization based on a statistical decision that is present in the speakers long term memory which characterizes the speakers 'grammar' and not the statistics themselves.

The 'decision process' aspect of statistical learning is the key, in our opinion, to understanding how and why we see 'complexity to simplicity' in emergent systems. The language learner begins his task awash in the complex surface data of spoken language. At first, the only available option is to do raw statistical analyses of the language. Once enough statistical analysis is completed, patterns will emerge and the learner can decide which patterns merit generalization and which patterns do not. The positing of generalizations from the initial statistical parse of the data now provides the learner with more information to work with. The posited generalization can be modified based on additional data and can be statistically analyzed as a source of new and ever more abstract generalizations. The overall effect of the cycle of analysis and decision making is the production of simpler and more parsimonious generalizations. The simplification of generalizations is a necessary condition for a statistical based learning algorithm to be useful. If there is no pattern in the statistical distribution of a set of data that can be generalized in a manner that is simpler than the distribution of the data itself, then there is no motive to posit that generalization. Thus, any memorized generalization must be simpler than the distribution of the data itself. This 'generalization condition' will hold over the entire cycle of statistical analysis, and one possible reason to stop the analysis is that there are no longer any patterns in the data that are simpler than the distribution of the data itself.

The relationship between complexity and simplicity in emergent systems

Our analysis of allomorphy in Tuvan in the previous section demonstrates how surface complexity can derived from the interaction of simple underlying generalizations. Our story about the acquisition of the 'plural' and 'adjective' suffixes in Tuvan demonstrate how a learner can extract simple generalizations from the complex surface data of spoken language. This demonstrates the 'complexity to simplicity' aspect of emergence as discussed by Alan Baker (this volume). Furthermore, there is an additional 'complexity to simplicity' aspect of this story that can be identified if we change our focus from acquisition to production for this set of Tuvan.

For acquisition, we assumed that the surface language was the source of complexity which needed to be analyzed into simpler generalizations. If we change our assumption from the underlying generalizations being 'simple' to the underlying generalizations being 'complex' we can see that the surface variation in the allomorphs under discussion show a 'simplicity' to them along the lines of the four seasons discussed by Alan Baker (this volume). Although, there are multiple distinct surface allomorphs for the 'plural' and 'adjective' morphemes, once we have the underlying generalizations they can all be identified easily and thus be considered 'simple' in a sense. Even though there is variation in the surface allomorphs, their distribution is predictable and stable analogous to the regular change of seasons.

The converse of the 'complexity to simplicity' aspect of emergence also discussed by Baker is also captured in our view of emergence. Although the data from Tuvan discussed in the previous section does not show this 'simplicity to complexity', the general analysis proposed for Tuvan is capable of producing this effect. An unstated aspect of the generalizations we posited for Tuvan is that they are 'transparent'. 'Transparent' in this case means that the generalizations posited are "surface true" in that all of the allomorphs in Tuvan discussed here fall out from the underlying generalizations. Because of this aspect of the data, there are multiple ways we could have constructed the original decision trees in (3) and (4). For expository reasons, we chose to put the 'consonant decision' leftmost in the decision trees but we could have put either of the other decisions (e.g. 'frontness decision', 'backness decision') leftmost in the tree and still derive the surface facts. In otherwords, the ordering of the generalizations in Tuvan does not matter. But this is only the tip of the proverbial iceberg of language complexity. All known human languages also contain many such generalizations that must be ordered in a specific sequential way.

The complex interaction of generalizations in phonology is referred to by the technical cover term 'opacity'. 'Opacity' can be defined as the situation where an essential generalization--one that a speaker must make to be competent user of the language--is not detectable in the surface data (e.g., in the speech patterns that are heard by the learner). The common source of this non-detectability is that with linear ordering, one generalization can either create or destroy the specific environment in which a different generalization is relevant, rendering one of them undetectable.

A classic example of 'opacity' can be found in the interaction of two generalizations, the first one called 'flapping' and the second 'Canadian Raising', found in some dialects of English (Idsardi 2006). 'Flapping' is a process where the sounds 't' and 'd', which are normally pronounced distinctly **(as in 'toe' vs. 'doe', or 'at' vs. 'add')** become merged into a single distinct sound in certain environments. This new sound is neither 't' nor 'd', but is what phoneticians call a ‘flap’ (we use the symbol ‘D’ for flaps in the data below), like the middle sound in the words 'utter' and 'udder'. (Note that highly literate speakers may resist the notion that 'bitter' and 'bidder' contain the same consonant in the middle, because they are thinking of the written form of the words. But acoustic analysis and perceptual tests confirm that these two sounds are pronounced identically in most English dialects).

‘Canadian Raising’ is a process where the diphthongs ‘ay’ (as in the word ‘bite’) and ‘aw’ (as in the second vowel in ‘about’) are altered to the diphthongs ‘ʌy’ and ‘ʌw’ respectively (the symbol ‘ʌ’ indicates a vowel that is pronounced higher in the oral cavity, and sounds quite different from ‘a’. A crucial aspect of Canadian Raising is that it only raises the vowel if it precedes a ‘voiceless sound’ which for the current example is ‘t’ but not ‘d’ or ‘D’. If this specific sound environement is not provided, the vowel raising process cannot occur.

The data in (6) show the interaction of ‘flapping’ and ‘Canadian Raising’ in two dialects of English. For full details of the importance of Canadian Raising and opacity see Idsardi (2006). The derivations (ordering of generalizations) for two dialects of English are presented below and the sequential order of generalizations is shown in descending order. Thus, for Dialect A, the Flapping generalization applies first and only alters the memorized form (enclosed in slashed brackets) if the environment is provided. After Flapping has had its chance to apply in Dialect A, then Canadian Raising can apply but only if the environment is provided.

(6) Interaction of Canadian Raising and Flapping (data from Chambers 1975:89-90 as cited in Idsardi 2006)

	‘writer’	‘rider’		‘writer’	‘rider’
Dialect A	/raytər/	/raydər/	Dialect B	/raytər/	/raydər/
Raising	rʌytər	------	Flapping	rayDər	rayDər
Flapping	rʌyDər	rayDər	Raising	-----	-----
Surface	rʌyDər	rayDər	Surface	rayDər	rayDər
	(opaque)				

The important interaction between Flapping and Canadian Raising can be seen in the different surface forms for Dialect A and Dialect B in (6). For Dialect A, Raising occurs first and thus the /t/ in ‘writer’ causes the change of /ay/ to /ʌy/ because /t/ is a ‘voiceless’ consonant and then Flapping occurs which changes both the /t/ in ‘writer’ and the /d/ in ‘rider’ to the flap [D]. The effect of this ordering of processes is that there is increased surface complexity in that we can not see why Canadian Raising should have applied to 'writer', because the needed environment for Canadian Raising is subsequently erased by Flapping. Dialect B applies Flapping first which changes both the /t/ and /d/ to the flap [D]. Consequently, when Canadian Raising attempts to apply, neither 'writer' or 'rider' have a 'voiceless' consonant and neither surface form shows the raised vowel.

Phonological opacity as presented in (6) is rife in human language, thus illustrating the 'simplicity to complexity' aspect of emergence in language structure. Whether phonological systems provide examples of 'simplicity to complexity' or 'complexity to simplicity' is then understood by whether the generalizations about the languages sound patterns interact in opaque ways when ordered.

A final note about opacity is the robust observation that it is present in all human languages and thus does not appear to cause a problem for language acquisition. This observation is crucial because it further supports our thesis that language learners posit abstract generalizations about the data they to which they are exposed. Because these generalizations must be simpler than the data itself, and are often decomposed into still simpler generalizations, we should not be surprised when the generalizations are 'put back together' in a particular order and surface complexity arises. Thus, since we propose that language learners are 'emergonauts' they are not surprised by the surface complexity of spoken language.

Human language and emergence

We hope our brief sketch of the acquisition of complex sound patterns has been helpful in exhibiting how human language is an emergent system. We understand human language and emergence to exist in a symbiotic relationship. Thus, the study of the structure of human language provides excellent naturalistic data from an emergent system. Further, the study of emergent systems will lead to deeper understanding of the structure of human language. The most beneficial aspect of our sketch in our opinion is the identification of specific and general questions about both human language and emergence that should be pursued further. One thing that must be understood about this chapter is that we are assuming that the required statistical analysis mechanisms and decisions that underlie the discovery of the relevant decision trees are present but do not yet know the full details of these mechanisms. We believe that the questions of what statistical analysis is present in language acquisition, what representations this analysis occurs on and what the decision processes are to posit generalizations are some of the most important questions that must be addressed, not only for language but for the cognitive sciences as a whole (). We hope this chapter brings a clarity as to why these are important questions and suggest useful ways to pursue them.

Our final goal of this chapter is to suggest that many (if not all) of the questions posed by this chapter directly benefit from the other work on emergence presented in this volume.

Only by understanding what the contributions on emergence from computer science, philosophy, neurobiology, economics and many other disciplines can we achieve the best understanding of human language as emergence.

References

Anderson, Gregory D. S. and K. David Harrison. (1999) *Tyvan*. München, Germany: Lincom Europa.

Aslin, Richard N., Jenny R. Saffran and Elissa L. Newport (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science* 9 (4): 321-324. [July].

Chambers, J. K. (1975). *Canadian English: Origins and Structures*. Toronto: Methuen.

Idsardi, William. (2006) Canadian Raising, Opacity and Rephonemicization. Ms. Univeristy of Maryland, College Park.