

# Parts and wholes: Patterns of relatedness in complex morphological systems and why they matter

Farrell Ackerman  
University of California at San Diego  
[fackerman@ucsd.edu](mailto:fackerman@ucsd.edu)

Robert Malouf  
San Diego State University  
[rmalouf@mail.sdsu.edu](mailto:rmalouf@mail.sdsu.edu)

The whole has value only through its parts, and the parts have value only by virtue of their place in the whole.

(Saussure 1959:128)

“...we cannot but conclude that linguistic form may and should be studied as types of patterning, apart from the associated functions. Sapir 1921:60

## 1 Introduction

Recent work on morphological learning and morphological theory has largely neglected an important research problem, which we will refer to as the *Paradigm Cell Filling problem*:

*Paradigm Cell Filling problem*: What licenses reliable inferences about the surface wordforms for the inflectional (and derivational) families of wordforms associated with (classes of) lexemes.

That is, given exposure to a novel inflected wordform, what are all the other wordforms in its inflectional (and derivational) families? How to predict the correct shape of words on the basis of limited experience with “similar” words is only a difficult problem because languages may depart from simple content/form mappings, some quite dramatically: in the simple case, of course, mappings would be automatic and transparent. But since they generally are not, with some languages containing numerous conjugation and declensions classes, the following questions arise:

- a. How are such complex systems organized?
- b. What role might this organization play with respect to licensing inferences concerning paradigm cell filling?

- c. What relation does this organization and the possibility for inferences based on surface patterns have concerning the learnability of complex systems?

These questions concerning the relation of parts to wholes, i.e., constitutive elements to the whole word configuration they appear in and whole words to the paradigms they participate in, as well as the essentially pattern-based nature of languages, represented here by words and paradigms, explore the intuitions evident in the twin themes of the epigrams above. In order to address these issues we will explore how Uralic languages (Finnish, Mordvin, Estonian, and particularly Tundra Nenets) provide fertile ground for identifying the nature of the challenges posed by the Paradigm Cell Filling Problem as well as assist in providing clues for possible solutions. We begin in Section 2 by grounding the present approach within the landscape of competing morphological theories. We identify two basic approaches which we refer to as *syntagmatic* and *compositional* versus *paradigmatic* and *configurational*, delineating their basic assumptions and suggesting that they identify different analytic objects with crucially different consequences for the nature of language inquiry. In Section 3, we provide the basic elements of our *paradigmatic/configurational* approach, defining the Paradigm Cell Filling Problems in its terms. Section 4 introduces the patterns of nominal inflection from Tundra Nenets that represent our case study for addressing the Paradigm Cell Filling Problem, while in Section 5 we compare two surface oriented approaches to morphological organization: we contrast what will refer to as an *asymmetric* approach word relatedness, typified by Albright (2002) with a *symmetric* approach (our own). We will test the predictions of these alternatives by examining Tundra Nenets declension classes and we will offer a provisional set of results about paradigm organization for this language.<sup>1</sup> We summarize our conclusions and speculate on their ramifications in Section 6.

## 1 The Morphological Landscape

A widespread approach to morphology, adapting the tradition of American structuralists and certain ideas of Europeans such as Badouin de Courtenay (Anderson 1985, Hockett 1987, Matthews 2001, among others), focuses on the small meaningful pieces which can be composed into complex words: this approach can be characterized as *syntagmatic*, because it emphasizes the linear combination of constitutive elements, and *compositional*, because it endeavors to derive the meaning of the whole word from the meanings of its identifiable parts. This so-called *morpheme-based* conception of the enterprise naturally leads to certain research questions, while precluding others. In particular, it leads to familiar efforts to identify small meaningful pieces (morphemes) as well as the rules (morphotactic and phonological) that yield the legal combinations of these entities evident as surface wordforms. It has also led to parsimony concerns regarding the minimal elements (either underlying or surface-based) and operations required to construct or build wordforms.<sup>2</sup> From this

---

<sup>1</sup> The fieldwork on Tundra Nenets was facilitated by a Hans Rausing Endangered Language Major Documentation Project Grant 2003–2006, for which Ackerman is extremely grateful.

<sup>2</sup> It should be noted that in practice these parsimony considerations have mostly been restricted to establishing the set of primitive representations, e.g., binary branching, and operations on tree-theoretic representation with uniform

perspective neither surface wordforms nor the systematic patterns of surface alternations that whole words participate in are construed as basic units of grammatical organization. Rather, the surface words of particular languages are useful to the degree that they provide insight into underlyingly invariant atoms of analysis and combinatoric operations that can account for (variable) surface expression. This all reflects a view whereby surface patterns of both words and networks of words (i.e., relations between surface alternants) are not considered to be proper objects of linguistic analysis, while the abstract elements and operations responsible for constructing these ephemera are. This, naturally, leads to questions concerning the psychological or biological basis of these constitutive elements and operations.<sup>3</sup>

There is another morphological tradition which is less familiar, though its historical antecedents are long and rich, that focuses on words and the ways in which related surface wordforms cohere into networks of wordforms. This approach can be characterized as *paradigmatic*, because it identifies (sets of) patterns that whole words participate in, and *configurational*, because, while the meaning of a wordform is not construed as a straightforward composition of individually meaningful parts, the meaning of the whole is associated with reliable arrangements of its constitutive elements. From this perspective the focus in morphology is shifted as in a Necker cube: instead of morphology being (solely) about the composition of complex wordforms from smaller pieces, it is about complex surface wordforms as representing types of configurations of elements and whole surface wordforms as elements in a network of related wordforms. As observed by Matthews 1991:204 “words are not merely wholes made up of parts, but are themselves construable as parts with respect to systems of forms in which they participate.” This emphasis on surface patterns of different sorts leads to a different set of research issues and questions. In particular, it becomes crucial to identify how complex words are organized into meaningful wholes without necessarily attributing meanings to identifiable parts, and how wordforms are organized into structured networks of conjugation and declension classes within inflectional and derivational families. In addition, it becomes natural to ask why the systems of organization cohere in the ways that they do, how such organization is learned, and whether the nature of the organization reflects learnability constraints, either specific to language or relevant in other learned domains as well. The paradigmatic/configurational perspective takes surface patterns seriously as entities that may facilitate learning, so the patterns represented by complex words and the patterns of organization among related words are, therefore, not the epiphenomenal result of representations and operations designed to produce individual words, as standardly assumed in syntagmatic and compositional approaches. In this *word and paradigm* perspective surface wordforms are not insightfully reduced to simple combinations of constitutive pieces but are better viewed as *recombinant gestalts* or configurations of recurrent partials (segmental or suprasegmental) that get distributed in principled ways among members of paradigms. As a consequence the domain of morphology can be seen as an instance of a complex system, and the analysis of language more broadly begins to look like it can benefit from methods in other fields which study complex systems. Within language this parallels

---

phrasal expansions. Reliance on this small inventory of representations and operations has led to increasingly abstract and some would suggest unparsimonious representations of rather simple surface expressions.

<sup>3</sup> See S. R. Anderson’s characterization of competing approaches to phonological analysis below.

what has been described independently within the domain of speech sounds by Oudeyer 2006:22 as the “systematic reuse”, and we would suggest *systemic reuse*, of phonological distinctions:

“all languages have repertoires of gestures and combinations of gestures which are small in relation to the repertoires of syllables, and whose elements are systematically reused to make syllables.”

Likewise in morphology, as exemplified in detail below, the basic elements of words are used again and again in different configurations. Since configurations convey different meanings, this diminishes the need for bi-unique relations between forms and functions, rendering their syntagmatic arrangement and composition less essential to the morphological enterprise.

The primary focus in the paradigmatic/configurational approach on surface words and their systematic alternants rather than on the identification of invariants responsible for deriving them recalls S. R. Anderson’s insightful overview of Badouin de Courtney’s and Kruszewski’s theory of alternations in phonology and morphology. Describing the evolving conception of the phoneme in the works of these two linguists, he argues 1985:68 that:

It is worthwhile to notice, however, that the issue of such an invariant element arises most directly as a consequence of the need to deal with the systematic *variance* represented by the alternations. It is this systematic variation, with its fundamentally relational character, that language presents us most directly. One way to organize this variation is to hypothesize underlying invariant units – indeed, judging from the history of the discipline, this is the most natural way for linguists to conceptualize such relations – but it should be borne in mind that this is not the only way to do, or even the most transparent... for example, Saussure seems to have held a view of the phenomenon of variance and alternation that was much close to an immediate account of the relations in question than to an account in terms of another kind of representation for linguistic forms, one given in terms of hypostatized invariants.

Crucially, the surface alternants were interpreted as participating in associative or paradigmatic networks of relations, requiring recognition of (networks of) whole words, hence, this approach is exemplary of the paradigmatic/configurational perspective advanced here.

### ***1.1 Claims and consequences***

In this section we compare the syntagmatic/compositional approach with the paradigmatic/configurational approach in connection with a simple and familiar data set, specifically, *regular inflection internal to lexical compounding* in English. We do this as a cautionary tale since it is our belief that the perspective on analysis assumed in the syntagmatic/compositional approach is based on questionable assumptions which have been consequential for much current research. In particular, from a methodological perspective this approach appears amenable to the two pronged criticism articulated in Corning 2005 concerning biological research:

“the extreme reductionist argument that an understanding of the parts fully explain the whole leads to what C. F. A. Pantin called the ‘analytic fallacy’”. ... a whole also represents a distinct, irreducible level of causation that harnesses, constrains, and shapes lower parts and which may, in fact, determine their fate.” (Corning 2005:94)

“The constructionist [= flip side of reductionism FA & RM] hypothesis breaks down when confronted with the twin difficulties of scale and complexity. At each level of complexity entirely new properties appear. (Anderson 1972 cited in Corning 2005:138)

With this in mind we will suggest that the predominant attention paid in the literature to phenomena in languages like English and German, with relatively simple morphological systems, leads to an erroneous view of the toolkit needed to address natural language morphological systems in general. Hence, theoreticians should be more responsive to a fuller taxonomy of morphological systems, especially attending to those where the issues of scalability and complexity are most evident.

Since the widely reported results of Gordon (1985) and Gordon and Alegre (1996) it has been claimed that English speaking 3 and 5 year old children appear to be aware of the adult-like distinction in acceptability between *mice-killers* and *\*rats-killers*<sup>4</sup>: while irregular plurals can occur as left-members in lexical compounds (as in *mice-killers*), regular plurals cannot (*\*rats-killers*). The purported categoricity of the simple split in acceptability in compounds has led to extraordinary claims concerning the structure of the human language faculty and the nature of language learning (e.g., Pinker 2000, Clahsen and Almazan 2001, Clahsen, *et al.* 2003). In particular, the Dual Route Model (Pinker 2000) is designed to preclude the possibility of regular inflection occurring internal to lexical compounds. This is schematized in Figure 1:

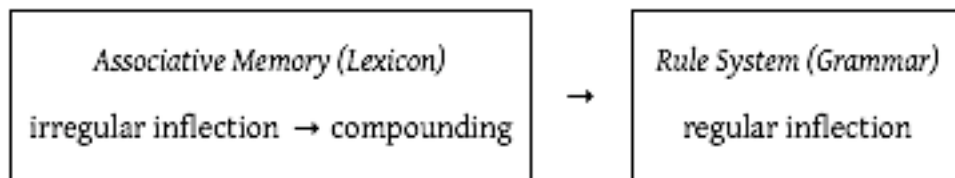


Figure 1 *Dual Route Model of Morphology*: No regular inflection internal to derivation

Since the plural of *mouse* is the irregular form *mice*, it is stored in the lexicon: it cannot be created by the application of a regular rule. Since it is stored, the compounding process in the lexicon has access to it and it can, as a result, combine with *killer* to yield *mice-killer*. In contrast, since *rat* is not associated with an irregular plural form, only the singular form is listed in the lexicon and only the singular form is available for lexical compounding. This precludes creation of the compound *rats-killer*, while allowing *rat-killer* to be formed. The regular plural formation operation can then apply to the compound and yield, if need be, *rat-killers*.

<sup>4</sup> See Ramscar (2005) on the (un)reliability of the judgments producing these data – they are claimed to be elicitation procedure artifacts.

To researchers in this tradition, this putative split and the earliness of behavioral (acquisition) evidence for it has suggested the possibility of biological constraints on the architecture of grammar, with consequences for synchronic grammar organization and language learning.<sup>5</sup> For example Clahsen 1999:1009<sup>6</sup> states that the “feeding relationships between plural inflection and compounding are determined by a grammatical ordering constraint. It is hard to see how children could learn this constraint directly from input data.” The implication here is that since such a constraint cannot be learned, it must be innate. This is a form of what Elman et al. refer to as *representational nativism*, the claim that the human species possesses an innate specification of substantive expectations and constraints that could yield the effect of e.g., regular inflection being absent internal to lexical compounding. In the present case, if there is a native and early predisposition against regular inflection internal to derivation one might expect that, all other things being equal, (1) grammars would not contain such constructions, since the (innate) language architecture itself prohibits it and (2) all children at their earliest stages would behave like English-speaking children in using only irregular inflection internal to compounds, whatever their target language.<sup>7</sup> We turn to each of these predictions in turn.

First, with respect to cross-linguistic synchronic grammars, it appears that there *is* regular inflection internal to derivation in synchronic systems.<sup>8</sup> In Sepečides-Romani (Cech 1995/1996:78), plural nominal forms can serve as bases for denominal verb derivation:

NOUN<sub>plural</sub> + *ndivola* → VERB<sub>get full of noun</sub>

Template for inchoative verb formation:

---

<sup>5</sup> This possibility is raised again in a recent series of experiments on English regular and irregular inflection internal to lexical compounds by I. Berent & S. Pinker (in press). They appropriately distinguish between the existence of a constraint and claims about the source of that constraint. They state their goal as follows: “Our interest here is whether the dislike of compounds like *rats-eater* is due to a constraint against regular plurals in compounds—an issue that is logically distinct from questions concerning the origins of that constraint.” We concur with the independence of these enterprises, but our own interest in this section is in how specific morphological assumptions appear to have encouraged particular beliefs about these origins or source of such a behavior. The biological basis of such constraints is baldly reaffirmed in a recent email correspondence between Dan Everett and Noam Chomsky as reported in the *New Yorker*: “UG is the true theory of the genetic component that underlies acquisition and use of language....[there is] no coherent alternative to UG.”

<sup>6</sup> See Scholz and Pullum (2005) for a carefully reasoned critical evaluation of reflexive appeals to innateness of the sort guiding Clahsen’s remarks.

<sup>7</sup> It is important to note that we have generalized the presumptive constraint from being specific to regular plurals to encompassing all regular inflection. We are, therefore, supposing that it seems more reasonable to assume a constraint that affects regular inflection irrespective of specific morphosyntactic properties than one which only targets e.g., regular number inflection or regular case inflection.

<sup>8</sup> It is generally recognized in the literature that compounding is a type of derivation. Therefore, these data are legitimate challenges, though they are not compounds.

On Cech’s account, there are some formations in which a plural interpretation of the nominal base is transparent:

- |     |                     |                      |                    |   |
|-----|---------------------|----------------------|--------------------|---|
|     | Singular            | Plural               |                    |   |
| (1) | <i>rukħ</i> ‘tree’  | <i>rukħa</i> ‘trees’ | O veš<br>the wood  | <i>rukħandivola</i><br>tree.PL.INCHOATIVE<br>‘The wood gets dense with trees.’    |
| (2) | <i>džuv</i> ‘louse’ | <i>džuva</i> ‘lice’  | O bala<br>the hair | <i>džuvandivola</i><br>louse.PL.INCHOATIVE<br>‘The hair is getting full of lice.’ |

As evident in (1) and (2), the suffix *-a* appears in plural forms of nouns. This plural form, in turn, appears in the derived verbs meaning ‘dense with trees’ and ‘dense with lice’, where the plural meaning is transparent. Similar formations apply to loanwords, as long as the semantics of the prospective predicate is acceptable. For example consider the following loanword from Turkish: (Cech 1995/1996:78)

- (3) *kurti* ‘worm’      *kurtja* ‘worms’      *kurtjandivola* ‘become full of worms’

As can be seen, the word ‘worm’ takes the plural form and this plural in turn used within the derived verb form. On any notion of default, this is the default nominal plural strategy: “Due to the steadily increasing number of loanwords incorporated in the masc. declension class, *-a* is the most abundant plural type within the noun declension in Sepečides-Romani.” (Cech 1995/1996:79).

Similarly, genitive nominals serve as a base for some types of verb formation in Tundra Nenets. In particular, this language possesses a process by which a nominal in the genitive plural serves as a base that hosts verbalizing suffixes. The verbs derived in this manner have either the meaning of possession, i.e., to possess X, where X is the entity denoted by the genitive plural, or to use in the capacity of X:<sup>9</sup>

NOM. SG.		GEN. PL		INFINITIVAL FORM OF DERIVED VERB
сађа	‘hat’	саби”	саби”ць	‘have a hat ~ use in the capacity of a hat’
ηум’	‘hay’	ηуђо”	ηуђо”ць	‘use instead of hay’

Table 1: Tundra Nenets Verbal Derivation

<sup>9</sup> Data from Kupriyanova et.al., 1985:139.

As is evident in Table 1, the verbal infinitival marker -*ць* is suffixed to the GENITIVE PL. form of the nominal. Thus, cross-linguistic morphological research reveals that synchronic language systems do exhibit regular inflection internal to derivation, so they had better be learnable!

Second, with respect to language acquisition, there *is* regular inflection internal to derivation in language acquisition. For this purpose it is sufficient to mention the results from a diary study and experiments in Finnish language morphological learning (Oulu dialect) by Vântillä and Ackerman 2000. In Finnish nominal inflection there are 15 case suffixes, two numbers (SG & PL). Compounds are right-headed (N N<sub>H</sub>), and non-heads tend to appear in various forms: they are uninflected (= NOMINATIVE<sup>10</sup>), GENITIVE-SG or GENITIVE-PL, but also can occur in various OBLIQUE cases. These alternatives are illustrated below: they represent the adult target that children must attain.

NOM.SG	<i>käsi-Ø-kauppa</i>	‘over-the-counter sales’
GEN.SG	<i>käde-<u>n</u>-puristus</i>	‘handshake’
GEN.PL	<i>käs-<u>ien</u>-hoitaja</i>	‘manicurist’,
ELAT.SG	<i>käde-<u>stä</u>-ennustaja</i>	‘hand-reader’
ADESS.SG	<i>käs-<u>illä</u>-seisonta</i>	‘handstand’

Table 2: Synchronic targets in Finnish

In the clearest instance of Finnish children’s use of regular inflection internal to compounding Vântillä and Ackerman 2000 found that in elicited production experiments for novel compounds 21 of 24 children aged 3;2-7;0 used genitive singular non-heads.<sup>11</sup> This is typified by the compound *lehmä-n-hoitaja* ‘cow-GEN.SG-keeper’, where the *-n* marking genitive singular is the sole allomorph for this case ending. As the only possible marker, of course, it must be interpreted as the regular or default marker. The children’s behavior with GEN.PL is somewhat more complex, but instructive. The GEN.PL has several allomorphs and therefore developing command over which form or forms are used with which nominal is more problematic than selecting the *-n* for GEN.SG. Despite this, GEN.PL allomorphs were also used internal to compounds when the adult form would contain GEN.PL, even though the child’s form did not correspond to the typical adult form. This can be seen in Table 3 where a boy 4;10 over-regularized the GEN.PL allomorph – *itten*:

Child:	<i>possu-<u>itten</u>+syöttäjä</i>	‘piggy-GEN.PL-feed-agent’
Adult:	<i>possu-<u>jen</u>+särkijä</i>	
Child:	<i>pullo-<u>itten</u>+särkijä</i>	‘bottle-GEN.PL-break-agent’
Adult:	<i>pullo-<u>jen</u>+särkijä</i>	

<sup>10</sup> Uninflected or NOMINATIVE marked non-heads represent approximately 75% of established lexical items. As seen below, however, inflected forms are quite productive.

<sup>11</sup> The same children used the same genitive singular form as independent words, i.e., not as members of compounds.

Table 3: Genitive Plural internal to compounds

Simplifying for present purposes, Finnish children’s early departures from the adult norm display forms reflecting regular case inflection internal to lexical compounds. This is particularly clear in their use of, e.g., the genitive singular, since the marker for this morphosyntactic category has no allomorphs, appearing only as the suffix *-n*, although it is also evident in the child’s reliance on a single form to mark GEN.PL as in Table 3. These markers must be analyzed as the regular, default form and, accordingly, children had better not be broadly prohibited from acquiring regular inflection internal to derivation. Thus, in Finnish, as in Sepečides-Romani, the adult grammar possesses a pattern apparently precluded by the relevant nativist constraint, while the evidence suggests that children too are not constrained in their development by the effects of such a presumptive constraint. Indeed, speculating on the research consequences of these results, Vântillä and Ackerman suggest that the existence of morphological systems with greater complexity, as in Finnish, may actually facilitate the learning of such systems by priming children to be sensitive to more distinctions and patterns earlier. If this is so, then learning about complex systems is a necessary task for identifying any generalizations likely to be true of morphology and its learnability.

Beyond these particular empirical problems with the proposed nativist constraint, the question arises from cross-linguistic research in morphology whether it is even intelligible to posit a gross constraint against inflection internal to derivation. As observed by several morphologists, all inflection is not the same, some seeming more derivational than others. In this connection, some researchers have proposed that there is a gradient between inflectional and derivational morphology (Booij 2002, 2005, Bybee 1985, among others). This can be schematized somewhat categorically as follows:

Root – derivation – inflection<sub>inherent</sub><sup>12</sup> - inflection<sub>contextual</sub>

While one might concede that the specific points raised above (the recognition of complex systems, apparent counterexamples in adult and developmental data) may prove problematic to some variant of the nativist proposal, an advocate of that line of inquiry might regard the cost of abandoning this type approach to language analysis as too great. After all, what about the standard scientific strategy of looking at simple systems, identifying in those systems the minimal atomic elements and combinatoric principles to produce the relevant target, and then scaling up when confronted with greater complexity? This strategy appears consistent with the goals and practice of such programs as Distributed Morphology (Embick and Noyer 2005, Embick 2007, among others) and Minimalist Distributed Morphology (Trommer 2003) where increasingly abstract representations have been posited. One consequence of these abstract tree representations is the claim that their psychological implausibility, i.e., their unlearnability with reference to surface stimuli, even more strongly implicates their innate status. But there are two real problems here. First, there is no guarantee (or even reasonable expectation) that understanding, say, English morphology in this

---

<sup>12</sup> Inherent refers to morphosyntactic categories such as semantic case and number, while contextual refers to morphosyntactic categories such as agreement and concord.

mode of analysis will shed any light on more complex systems such as Tundra Nenets (see below) or Chinantec (see Stump and Finkal this volume), hence the previous reference to problems of scalability and complexity. Second, and perhaps more worrisome, as observed in Braine 1992:79 about the far less abstract representations imputed to biology in Pinker's 1984 speculations about syntactic category development in children:

Within the syntactic theory there is clearly a difficult scientific problem about the origin of syntactic categories: how do we get from genes laid down at conception to syntactic categories manifest two-and-a-half to three years later? Merely labeling the categories as innate does not solve the problem; it just passes the problem to biology without considering how the biologist could ever solve it... While it is certainly not now reasonable to demand anything like a complete theory, it *is* reasonable to expect a promissory note, and at least a sketch of an argument as to how it might eventually be redeemed.

In some sense, the advocates of syntagmatic/compositional morphology appear to believe that understanding English and Dutch will be as consequential or illuminating as understanding *Drosophila manogaster* (the fruit fly) has been in developmental biology. In particular, a key insight informing modern developmental biology is that much of the evident diversity in the body shape and body plan systems of living creatures derives from demonstrable genomic similarities, and, furthermore, that the mechanisms responsible for this constrained variability, the genetic 'toolkit' guiding embryonic development, are largely the same across all higher organisms. The development of, e.g., eyes and limbs in insects are controlled by very similar genetic mechanisms, as are homologous structures in mammals. This means that studying the development of simple species, like the fruit fly, can shed considerable light on the development of more complex species, including humans.

However, at the beginning, it was by no means obvious that developmental models of simple creatures could tell us anything about more complex organisms:

For more than a century, biologists had assumed that different types of animals were constructed in completely different ways. The greater the disparity in animal form, the less (if anything) the development of two animals would have in common at the level of their genes. One of the architects of the Modern Synthesis, Ernst Mayr, had written that "the search for homologous genes is quite futile except in very close relatives." But contrary to the expectations of any biologist, most of the genes first identified as governing major aspects of fruit fly body design were found to have exact counterparts that did the same thing in most animals, including ourselves. (S. Carroll 2005:9)

Indeed, earlier work on the bacteria species *E. Coli* had turned out not to generalize to more complex organisms (Keller 2002). The discovery of a universal genetic toolkit was the breakthrough that has led to a revolution in our understanding of evolution, and Edward B. Lewis, Christiane Nüsslein-Volhard, and Eric F. Wieschaus, three important figures in this development, were awarded the Nobel Prize in 1995 for their work.

It may, of course, turn out that a similar situation obtains in language morphology, and that a thorough understanding of e.g., English or Dutch morphology in its simplicity may tell us all we need to know to understand about the organization of Finnish, Tundra Nenets, or Chinantec. Dissenting from this strategy, B. L. Whorf 1956:215 expresses concern about the sort of linguistic science that would be based on familiar Indo-European patterns, let alone the simplest of these, in neglect of those found in the native languages of the Americas:

To exclude the evidence which their languages offer as to what the human mind can do is like expecting botanists to study nothing but food plants and hothouse roses and then tell us what the plant world is like.

It is not a sufficient response in this connection to observe that morphological phenomena from numerous languages, some quite complex, have been analyzed employing the representations and assumptions developed for the simpler languages. Whorf's point, and ours, is that it is possible, even likely, that the explanatory toolkit appropriate for addressing surface diversity in linguistic morphological systems consists of elements and principles quite different than the reductive models postulated within mainstream generative approaches to grammar analysis guided by syntagmatic/compositional assumptions. We will only know this after we have compared more detailed and comprehensive descriptions of complex morphological systems and explored explicit analytic alternatives such as those based on paradigmatic/configurational assumptions.

Continuing the biological metaphor, we can characterize the mainstream syntagmatic/compositional theoretical gambit as representing, in effect, a *monogenic theory of inflection*. The monogenic theory of inflection, like biological proposals for monogenic disorders, posits a determinate relation between genetic representations and some outcome, here, inflection in grammar, i.e., there is some gene or collusion of genes responsible for such specific morphological phenomena as regular inflection internal to derivation. Criticizing this view of the relation between genes and behavior (genotype to phenotype) within biology Jablonka & Lamb 2005: 57 write:

It appears to invoke a view of the direct relation between genotype and behavioral expression characteristic of monogenic disorders. "In the case of simple monogenic disorders like sickle cell anemia, people with the defective genes always have the symptoms, whatever their conditions of life and whatever other genes they have. However, such simple monogenic diseases are not common: they make up less than 2 percent of all the diseases that are known to have a genetic component. For the remaining 98 percent of 'genetic' disorders, the presence or absence of the disease and its severity are influenced by many genes and the by the conditions in which a person develops and lives. Unfortunately, many people's understanding of the relation between genes and characters is based on the tiny minority of monogenic diseases. The popular view is that genes discretely and directly determine what a person looks like and how they behave.

The monogenic theory of inflection, however, seems to reflect a strategy advocated more broadly for the analysis of grammar and the human language faculty. This is evident in a recent article by

Anderson and Lightfoot 2006:381 in which they clarify their view of biology as it guides their linguistic theorizing:

Investigators have long sought for a very long time to understand how the functions of an organism follow from its species-specific biology, and that work was part of the “grand synthesis” of biology at the beginning of the twentieth century. Indeed, that was the importance of Gregor Mendel’s work on varieties pea-plants, and modern linguists whose work we describe often see themselves as doing a kind of Mendelian genetics, teasing out properties that seem not to be learnable from the environment and which must be provided in some fashion, and doing so by constructing poverty-of-stimulus problems in ways analogous to Mendel’s methods. (Anderson and Lightfoot reply to Everett)

In response to overly deterministic speculations about the relation between biology and behavior, West-Eberhard 2003:10 suggests a more dynamic synergy between genes and context in her comprehensive synthesis of biological research *Developmental Plasticity and Evolution*:

The naked ignorance of supposing that a genome could represent a blueprint for an organism is exposed by realizing that all gene expression depends on preexisting phenotypic structure and specific *conditions* as surely as upon specific genes. The phenotype is cohesive, but it is also eminently changeable.

In line with this more nuanced perspective from a field which excels in detailed description of variable expression analyzed with tools and methods of estimable rigor, we offer a speculative alternative for approaching morphological analysis and grammar more broadly.<sup>13</sup> This is what can be called the *epigenetic* theory of inflection. On such an approach, innately guided predispositions, not necessarily language specific, are responsible for the creation and maintenance of inflectional systems. The basic line of inquiry and its consequences are intimated in the following:

Behavioral development is the emergent product of a complex process of epigenesis, to which both genes and environments contribute, but neither genes nor the environment code for behavior directly. Both sides of the nature-nurture debate share the same erroneous assumption that the instructions for behavior are pre-existent in the genome or the environment and are imposed from without on the developing organism. Instead, genetic and environmental influences are inputs to a developmental process and their impact on behavioral outcome depends on their interactions with all of the components of that process. Consequently, it is misleading to speak of the genome as a “blueprint” or to think that genes “code” for behavior. Genes simply code for protein structure, and variations in the structure of a given protein, in a particular epigenetic context, may push behavioral outcomes in one direction or another. So, genetic and environmental factors are best

---

<sup>13</sup> This perspective is developed in terms of construction-theoretic approaches to grammar analysis in both morphology and syntax in Ackerman, Nikolaeva, and Malouf (to appear). This represents an extended and detailed case study of pronominal relative clause construction in numerous languages of Eurasia.

conceptualized as acting as risk (or protective) factors in the development of individual differences in behavior; their effects are probabilistic rather deterministic. (Pennington 2001:496)<sup>14</sup>

This epigenetic perspective, we believe, is essentially compatible with the paradigmatic/configurational view of complex words as (types of) patterned arrangements of elements and of words as representing patterns of surface alternants defining systems of (sub)paradigms. On this view, contra the monogenic approach, there is no need to genetically specify particular aspects of e.g., inflection; instead they arise systemically and systematically in much the way that surface complexity arises from simple interactions in biological systems:

Relatively little needs to be coded at the behavioral level and the information required for action by the individual is often local rather than global. In place of explicitly coding for a pattern by means of a blueprint or recipe, self-organized pattern formation relies on positive feedback, negative feedback, and a dynamic system involving large numbers of actions and interactions. With such self-organization, environmental randomness can act as the ‘imagination of the system’, the raw material from which structures arise. Fluctuations can act as seeds from which patterns and structures are nucleated and grow. The precise patterns that emerge are often the result of negative feedback provided by these random features of environment and the physical constraints they impose, not by behaviors explicitly coded within the individual’s genome. (Camazine, *et al.* 2001:26)

This kind of epigenetic approach to understanding language morphology has antecedents in efforts to understand various other aspects of human behavior from a developmental perspective. A particularly clear articulation of this approach can be found in L. Smith & E. Thelen 2003:347:

The major problem for a theory of development is to explain how to get something more from something less. At multiple levels of analysis at multiple timescales, many components open to influence from the external world interact and in so doing yield coherent higher-order behavioral forms that then feedback on the system, and change that system. In human development, every neural event, every reach, every smile and every social encounter sets the stage for the next and the real time causal force behind change. If this is so, then we will gain a deeper understanding of development by studying multicausality, nested timescales and self-organization.

In sum, a paradigmatic/configurational approach to language morphology, which takes as its foundational elements the pattern-based nature of (classes of) complex words

---

14 The references to biology are intended to contextualize morphological analysis within the “EvoDevo” spirit of J. Blevins’ (2004) explorations in phonology, which itself, of course, recalls the speculations concerning the interaction between diachrony and synchrony in Badouin de Courtenay and Kruszewski (see S. R. Anderson 1985 and Durand and Laks 2002.)

as well as the pattern-based nature of the systems of relatedness that words participate in, finds useful antecedents in the methods and assumptions applied to the analysis of complex systems in both developmental biology and other areas of human development. Recognizing that the syntagmatic/compositional approach has led to certain assumptions about the relation between language and biology, we have looked at recent trends in developmental biology and speculate that they offer linguists a way of looking at language that is compatible with the views independently arrived at on linguistic grounds in the paradigmatic/configurational approach. In the remainder of this article we direct ourselves to specifying how this linguistic approach addresses the Paradigm Cell Filling Problem.

## 2 The basic analytic problem

Having broadly delineated the landscape of morphological alternatives and how they align naturally with different research enterprises, we turn to the Paradigm Cell Filling Problem. In order to do this we begin with an overview of what it means to be a paradigmatic/configurational approach to morphology.

Following Trosterud (2004:54), we will assume that:

Morphology is essentially about relations between whole words (*paradigmatics*), not about pieces that make up single words (*syntagmatics*): “wordforms are signs, parts-of-wordforms are not.

As a consequence, it is crucial to specify a working definition for the construct paradigm:

The PARADIGM of lexeme *L* is the set of pairs of morphosyntactic [grammatical] words and wordforms that realize *L*. (Trosterud 2004:15)

On this construal a paradigm is a (multidimensional) matrix of morphosyntactic properties whose cells define legal combinations of features for lexemes of a specified category (i.e., *morphosyntactic/grammatical words*) and whose occupants are surface wordforms of lexemes.<sup>15</sup> That is, following Ackerman and Webelhuth 1998, Stump 2001, 2003, 2005, Ackerman and Stump 2004, among others, a paradigm is the domain in which content-theoretic (meaningful) notions are set in a principled correspondence with their form-theoretic surface expression. Sometimes we find a one-to-one mapping between content and form, as in so-called ‘agglutinating’ languages. But, we also regularly find more complex many-to-many relations between morphosyntactic property sets and wordforms<sup>16</sup>, with the same formal pieces used for different functions in different wordforms (homonymy/syncretism). For example, in Tundra Nenets, the same members of a suffix set can be used with different lexical categories, sometimes serving essentially the same function, and sometimes serving different functions.<sup>17,18</sup>

---

<sup>15</sup> Matthews 1991; Stump 2001, 2002; Ackerman & Stump 2003, Trosterud 2004, among others.

<sup>16</sup> See Trosterud (2004/to appear) for an insightful exploration of this issue within Uralic agreement systems from a word-based morphological perspective, especially focusing on why syncretism occurs and why it occurs where it does.

<sup>17</sup> This discussion follows the presentation in Salminen (1997:.96; 103; 126).

	N	V
Suffix set I	Predicative	Subjective
Suffix set II	Possessive	Objective

Table 4: Suffix homonymy in Tundra Nenets

As schematized in Table , markers from Suffix Set I can appear on both nouns and verbs, and the inflected word functions as the predicate of the clause. In either case, the set I markers reflect person and number properties of the clausal subject. While markers from Suffix Set II similarly occur with either nouns or verbs, their function differs within each class: they reflect person/number properties of the possessor when they appear with N, but number properties of clausal objects when they appear with (transitive) verbs. There is, in other words, a configurational dynamic whereby the same elements in different combinations are associated with different meanings. Inflected words, accordingly, are best construed as *recombinant gestalts*, rather than simple (or even complex) combinations of bi-unique content-form mappings (i.e., morphemes).<sup>19</sup> This perspective on complex words is intimated in Saussure's *Cours* 1966:128 in his discussion of *associative* (= paradigmatic) relations:

A unit like *painful* decomposes into two subunits (*pain-ful*), both these units are not two independent parts that are simply lumped together (*pain + ful*). The unit is a product, a combination of two interdependent elements that acquire value only through their reciprocal action in a higher unit (*pain × ful*). The suffix is non-existent when considered independently; what gives it a place in the language is series of common terms like *delight-ful*, *fright-ful*, etc.... The whole has value only through its parts, and the parts have value by virtue of their place in the whole.

Accordingly, while we are often able to isolate pieces of complex, it is the configurations in which these pieces occur and the relation of these configuration to other similar configurations that are the loci of meanings relevant in morphology. This property becomes even more evident by considering the structure of Tundra Nenets verbs as schematized in Salminen 1997:

---

<sup>18</sup> While predicate nominals and adjectives in Tundra Nenets host markers from Suffix set I, they differ from the verbal predicates which host these suffixes in exhibiting nominal stem formation rather than verbal stem formation and the inability to host future markers, and in their manner of clausal negation. All of these argue that two different lexical categories host markers from Suffix set I, and that there is no N-to-V conversion operation.

<sup>19</sup> See Gurevitch 2006 for an analysis along these lines for Georgian.

CONJUGATION	NUMBER OF OBJECT	MORPHOLOGICAL SUBSTEM	SUFFIX SET
<i>subjective</i>		general finite stem	I
	<i>sg</i>	(modal substem)	II
<i>objective</i>	<i>du</i>	dual object (modal) substem	III
	<i>pl</i>	special finite stem	
<i>reflexive</i>		(special modal substem)	IV

Table 5: Exponence of Tundra Nenets verbal forms as a function of conjugation

The general picture from Table is that there is little one-to-one correspondence between any cells across columns. That is, if one considers the column containing the general finite stem, exemplified in (4), we see that the stem serves as base for both subjective (4a) and objective conjugation (4b), as well as for the singular number for objects as marked by suffix set II. Likewise, as exemplified in (5), the dual object (modal) substem (5a) hosts members of suffix set III, but this set also serves to mark plural objects with the special finite stem (5b). Finally, the special finite stem is not restricted to plural object conjugation, since, as shown in (5c), it is associated with the reflexive conjugation and this conjugation's characteristic distinctive use of suffix set IV.

General finite stem:

- (4) a. subjective: *tontaø-d<sup>0</sup>m*  
 cover.I (=1 sg.)  
 'I cover (something)'
- b. objective sg.: *tontaø-w<sup>0</sup>*  
 cover.II (=1 sg/sg)  
 'I cover it'

*Dual obj. Stem:*

- (5) a. objective du. *tonta-gax<sup>0</sup>yu-n<sup>0</sup>*  
 cover.dual.III (=1 sg/du.)  
 'I cover them (two).'

*Special finite stem:*

- |                  |   |
|------------------|---|
| b. objective pl. | <i>tonteyø-n<sup>0</sup></i><br>cover.III (=1sg/pl.)<br>'I cover them (plural)' |
| c. reflexive     | <i>tonteyø-w<sup>0</sup>q</i><br>cover.IV (=1sg)<br>'I got covered'             |

In sum, it is the pattern of all the individual elements in their specific combinations that are the realizations of the relevant lexemic and morphosyntactic content associated with words, rather than the simple sum of uniquely meaningful pieces that is important.

This recombinant quality of morphology, in conjunction with the biological analogy of morphology as a complex system, clearly raises different questions than those derived from a syntagmatic/compositional perspective. In particular, it is crucial to consider what one can posit about natural language morphological systems and “optimal” organization. Though this is not an issue to be directly addressed here, it is evident that before one can say much it would be best to understand the extraordinary complexifications of morphological systems such as<sup>20</sup>

- a. the wild profusion of nominal declension classes in Estonian relative to earlier Finnic and Uralic nominal declension (see Blevins 2005, 2006)
- b. the extraordinary and unprecedented articulation of agreement dimensions and attendant surface exponence in Mordvin verbal conjugation relative to all other Uralic systems.

Restricting ourselves to the latter, it is sufficient to identify the following Uralic object agreement marking patterns expressing relations between the person/number of SUBJ and person/number of OBJ (Keresztes 1999). In Hungarian object agreement reflects a relation between all three persons in the SUBJ and 3<sup>rd</sup> person singular OBJs, with a special marker indicating a relation between 1<sup>st</sup> singular SUBJ and 2<sup>nd</sup> person OBJ. In Vogul, Ostyak, and Tundra Nenets these relations are somewhat more articulated: all person/number combinations for SUBJ are associated with distinctive marking patterns for all three number distinctions in OBJs. Finally, the most complex system is the innovative one found in Mordvin where, with a fair bit syncretism, there are correspondences between all person/number properties of SUBJs and OBJs. These systems are schematized in Figure 2, and an illustrative subparadigm of Erza Mordvin object agreement is presented in Table .

---

<sup>20</sup> We mention only Uralic here, but the reader should also consult Gurevitch (2006) for a careful examination of Georgian in this connection.

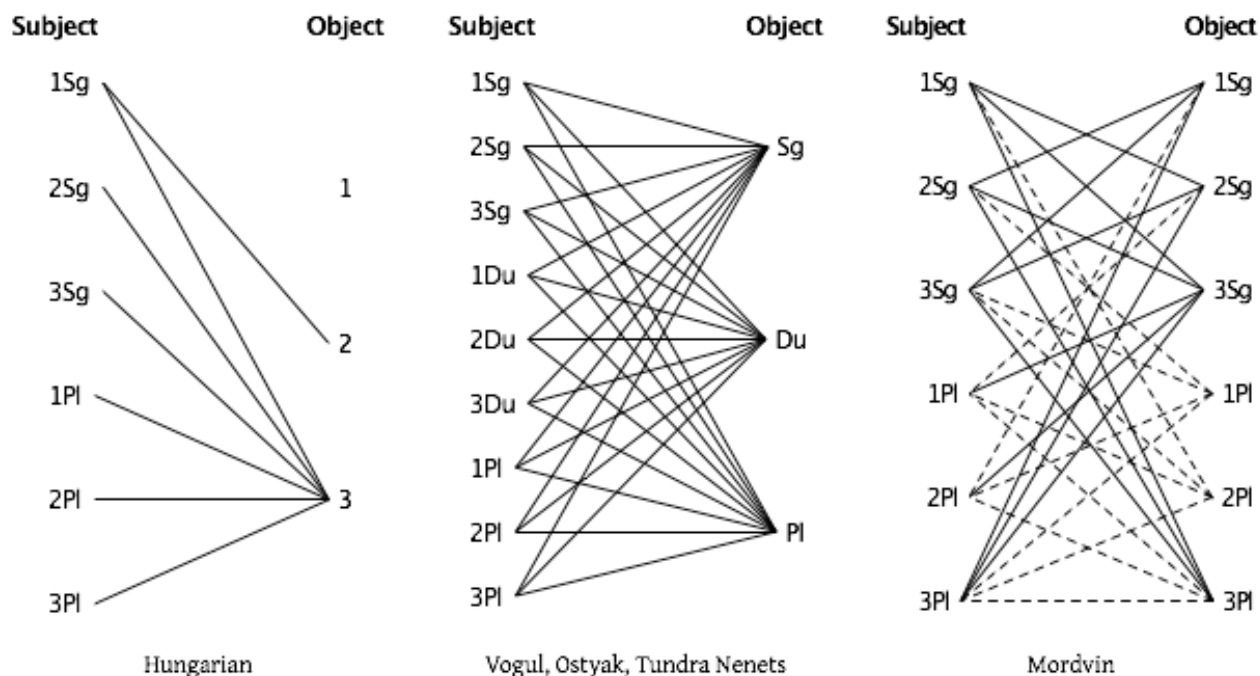


Figure 2: Uralic subject and object marking

Subj/Obj → ↓	3sg	3pl
Sg 1	<i>pala-sa</i> 'I love him/her'	<i>pala-siñ</i>
2	<i>pala-sak</i> 'you love him/her'	<i>pala-si'</i>
3	<i>pala-si</i>	<i>pala-siñže</i>
Pl 1	<i>pala-siñek</i>	<i>pala-siñek</i>
2	<i>pala-sink</i>	<i>pala-sink</i>
3	<i>pala-siž</i>	<i>pala-siž</i>

Subj/Obj → ↓	1sg	1pl
Sg 2	<i>pala-samak</i> 'you love me'	<i>pala-samiž</i>
3	<i>pala-samam</i> 'he loves me'...	<i>pala-samiž</i>
Pl 2	<i>pala-samiž</i>	<i>pala-samiž</i>
3	<i>pala-samiž</i>	<i>pala-samiž</i>

Table 6: Erza Mordvin: *palams* 'to love' (following Keresztes 1990:46)

The development of the Mordvin system from the much simpler antecedent system in Proto-Uralic, and its divergence from its synchronic congeners, presumably presents the sorts of chal-

lenges that an adequate theory of morphology should attempt to address: namely, what about the system of these languages licenses these patterns and may lead to the observed increases in apparent complexity? In this connection, it is worth noting that there is systemic cost associated with the Mordvin innovation. Specifically, the objective agreement paradigms in Hungarian, Vogul, Ostyak, and Tundra Nenets are all identical to some case/number paradigm for nominal possessives in those languages (e.g., set II above for Tundra Nenets), whereas we do not find this correspondence between the objective agreement paradigm and the possessive paradigm in Mordvin—the parallelism between these paradigms is lost. When looked at systemically, then, there are costs to morphological reorganization, recalling, we suggest, the sorts of multidimensional factors typically responsible (re)organization of biological systems. Finally, why should there be such evident stability in complex systems? Each of these agreement systems, displaying variable degrees of complexity, have proved to be quite stable over time. This resilience is evident in other domains and other languages, as observed by Gurevitch (2006:72) for intricate inflectional system of Georgian. She cites A. Harris (1985):

Yet upon closer examination, Georgian morphosyntax is highly regular and some of its key features have been stable since the earliest attestations of Old Georgian in the 5<sup>th</sup> century (Harris 1985).

For the moment, this can all be summarized by simply questioning whether there is any presently useful sense in which these Uralic verbal marking systems are more (or less) optimal than their historically antecedent systems, which arguably may not have had object agreement at all, or the synchronic systems of related languages. We cannot, of course, answer this or related questions here. It is simply our goal to illustrate the type of questions which naturally arise given a paradigmatic/configurational morphological perspective. We conclude this section by hypothesizing that there is simply too much specificity and variability in inflectional systems and too much of a role for systemic interactions among wordforms for much to be meaningfully posited as “innate”. We are left, then, with two basic questions. What are the principles of organization for these complex systems, and what learning mechanisms might facilitate their acquisition? For this we turn to the Paradigm Cell Filling Problem in earnest.

For languages with several distinct inflectional classes, resolving which class a word belongs to is sufficient to complete its paradigm, but the means for resolving class membership is sometimes complicated.<sup>21</sup> From a lexicographer’s perspective, one can identify one or more *principal parts*, paradigm cells whose shape is diagnostic of the inflectional class of a lexeme. These are the minimal forms which must be listed in a dictionary or memorized by a student to allow production of the complete paradigm. This reduces the Paradigm Cell Filling problem to the Inflectional Class Assignment problem – once we know which inflectional class a word is a member of, generating the complete paradigm is straightforward.

However, the actual problem confronted by language learners is quite different, since they have no guarantee that they will encounter the diagnostic wordforms necessary for resolving a

---

<sup>21</sup> For a transparent exposition of this issue and a novel solution see Stump & Finkal (this volume).

word’s inflectional class. Even in the simplest situation, where all members of a paradigm are predictable on the basis of a single form, there is no reason to expect that the language learner will have encountered that form for every word. What may be sufficient for the abstract purpose of generating full paradigms in a dictionary may turn out not to be useful for learners in developing command of their morphological systems. On the other hand, in many cases it is possible to predict one member of a word’s paradigm from another even though neither form uniquely identifies the word’s inflectional class. From a language user’s perspective, the Paradigm Cell Filling problem is quite distinct from the Inflectional Class Assignment problem, and reducing the former to the latter is not a useful strategy.

It will be useful to here to consider a (relatively) simple example. Finnish (Uralic) nouns are marked for case (NOM, GEN, PART, ...) and number (SG, PL).<sup>22</sup> A representative sample of Finnish declension classes appears in Table :

Class	NOM.SG	GEN.SG	PART.SG	PART.PL	INESS.PL	
4	<i>lasi</i>	<i>lasin</i>	<i>lasia</i>	<i>laseja</i>	<i>laseissa</i>	‘glass’
9	<i>nalle</i>	<i>nallen</i>	<i>nallea</i>	<i>nalleja</i>	<i>nalleissa</i>	‘teddy’
8	<i>ovi</i>	<i>oven</i>	<i>ovea</i>	<i>ovia</i>	<i>ovissa</i>	‘door’
32	<i>kuusi</i>	<i>kuusen</i>	<i>kuusta</i>	<i>kuusia</i>	<i>kuusissa</i>	‘six’
10	<i>kuusi</i>	<i>kuuden</i>	<i>kuutta</i>	<i>kuusia</i>	<i>kuusissa</i>	‘spruce’

Table 7: Schematic partial paradigm for Finnish nominal declension classes<sup>23</sup>

The shaded cells in Table are diagnostic in that they uniquely resolve class assignment for a lexeme, while plain wordforms do not. Thus, we find the following generalizations, whereby predicative wordforms identify correct class assignment. Given the stimulus *tuohtha* ‘birchbark (PART.SG)’, there is correct assignment to class 32 based on the analogical proportion  $kuusta : tuohtha :: kuusi : TUOHI$ . In contrast, where the stimulus is a non-diagnostic wordform, correct class assignment is underdetermined. Thus, the stimulus *nuken* ‘puppet (GEN.SG)’ could be assigned either to class 9 or class 8, based on the competing analogical proportions  $nallen : nuken :: nalle : NUKKE$  versus  $oven : nuken :: ovi : NUKKI$ . However, if the stimuli comprise the pair *nuken* ‘puppet (GEN.SG)’ and *nukkeja* ‘puppet (PART.PL)’, then there correct assignment of this word to class 9. That is, the conjunction of the forms in these cells is diagnostic. This, of course, stands in stark contrast to proposals that endeavor isolate a single a single form (either a stem, an underlying wordform, or a surface wordform) as sufficient to generate all inflected variants of a word.

<sup>22</sup> We follow the basic representations and lines of argumentation in Paunonen 1976, Thymé 1993, and Thymé, Ackerman, and Elman 1994. Standard accounts posit that Finnish possesses 15 nominal cases and two number distinctions.

<sup>23</sup> The numbers in the *Class* column refer to declension classes as presented in the *Soome-eesti sõnaraamat* (Finnish-Estonian Dictionary) Kalju Pihel & Arno Pikamäe (eds.) 1999. Tallin: Valgus.

This is (essentially) Stump and Finkel's notion of *dynamic* principal parts, contrasting with *static* and *adaptive* analyses (see this volume). In fact, there are many equally good alternative sets of principal parts for Finnish, and many more solutions that are almost as good. For example, the form *kuusia* 'six (part.pl.)' is not sufficient to identify the word's inflectional class or to predict all the other members of the paradigm. It is sufficient, though, to correctly predict that the nominal singular is *kuusi* and the inessive plural is *kuusissa* – filling these cells in the paradigm does not require resolving the word's inflectional class. Indeed, we speculate that this is a common feature of complex morphological systems (reminiscent of *resilience* in biological systems). Even though there may be a few very hard cases, in general most cells in the paradigm of most words are predictable from most other cells. This state of affairs is both surprising and inexplicable if speakers are using the lexicographer's strategy for paradigm completion.

This can be construed as a general hypothesis concerning (sub)paradigm organization according to which identifiable patterns of relatedness among wordforms in a paradigm facilitate paradigm cell filling. Moreover, related wordforms are partitioned into (sub)paradigms with their own small systems of relatedness among forms. What Finnish (sub)paradigms share, on this analysis, are recurring formal elements. For example, *lasi* occurs in NOM. SG., GEN. SG., and PART. SG., while *lase* occurs in PART. PL. and INESS. PL.<sup>24</sup> The force of this generalization can be formulated as follows:

Paradigm Cell Filling problem (general formulation):

Given a lexeme *L* associated with a set of morphosyntactic properties and expressed by a surface wordform, what are the surface wordforms for all other possible morphosyntactic property sets of *L*, i.e. what is the complete paradigm of surface wordforms for *L*?

Thus, paradigm cell filling concerns the licensing of reliable inferences about the surface wordforms for the inflectional (and derivational) families of wordforms associated with (classes of) lexemes. In other words, given a novel inflected word form, what are all the other wordforms in its inflectional (and derivational) families?

There are certain aspects of this generalization that are important to highlight. The first concerns the nature of the claim about what precisely forms the basis for prediction within and across classes. Our hypothesis is that this implicates the significance of surface words. In particular, the inferences are based on surface words and patterns of relatedness among surface words<sup>25</sup>. We interpret the surface word in accordance with the proposals such as Ackerman and Stump where a lexeme and an associated set of morphosyntactic properties can receive either synthetic or multiword, periphrastic, i.e. exponence. In either case, these surface expressions are inter-

---

<sup>24</sup> We restrict our focus here to "stems" and their reuses across paradigm rather than to "markers" and their reuses as the formal elements that recur, i.e., stem versus formative syncretism.

<sup>25</sup> Albright & Hayes 2003, Albright 2002, Anderson 1992, Aronoff 1993, Blevins 2006, Bochner 1993, Booij 2005, Bybee 1985, Kirby 2006, Matthews 1991, Neuvel and Fulop 2002, Skousen 1989, Trosterud 2004/to appear, among many others).

puted as the occupants of cells in paradigms.<sup>26</sup> The postulation of the importance of surface words leads to a correlative hypothesis concerning the relation among families of surface words. We express this hypothesis as follows:

*Patterns in the word system*

Patterns of relatedness between wordforms partition morphosyntactic feature combinations into (sub)paradigms which cohere with respect to the recurrence of “formatives” constitutive of wordforms, i.e., configurations of “recurrent partials” such as segments, tones, etc.

Thus for any given language one must inquire about (1) What the (sub)patterns of (inter)predictability are and (2) what elements play a role in the (inter)predictability among patterns. These are the issues we turn to in next section where we examine Tundra Nenets nominal inflection.

### 3 Patterns of predictability in Tundra Nenets

Following the outline of issues raised previously for Finnish, the present section provides a preliminary case study of nominal inflection in Tundra Nenets (Samoyed branch of Uralic). The basic question is this: Given any Tundra Nenets inflected nominal wordform, what are the remaining 209 forms of this lexeme for the allowable morphosyntactic feature property combinations CASE { nom, acc, gen, dat, loc, abl, pro }, NUMBER {singular, dual, plural}, POSSESSOR {3 pers. x 3 num.}? The problem can be schematized as in (6a) and (6b). Specifically, given exposure to a stimulus such as that in (6a), the nominal *nganuqmana* ‘boat (plural prosecutive)’, what leads to the inference that its nominative singular form is the target *ngano*? In contrast, if confronted with the plural prosecutive of the nominal *wingoqmana* ‘tundra (plural prosecutive)’, what leads to the inference that its nominative singular is the target *wih*?

(6) a.	Stimulus:	Target	vs.	b) Stimulus	Target
	<i>nganuqmana</i>	<i>ngano</i>		<i>wingoqmana</i>	<i>wih</i>
	boat.PL.PROS	boat.SG.NOM		tundra.PL.PROS	tundra.SG.NOM

In line with the hypotheses stated in the previous section, what we must do is identify the patterns of interpredictability for a subset of Tundra Nenets nominal declensions within and across (sub)paradigms. This entails stating the principles of patternment within and across stem types.

For the Tundra Nenets *absolute* declension (i.e., non-possessive non-predicative nominals), lexical categories are divisible into the following gross stem-type classification (ignoring the relevance of syllabicity, see Salminen (1997, 1998) for careful exposition of Types and see VI below for use of these classes):

---

<sup>26</sup> This does not diminish the importance of phonology in the syntagmatic composition of whole word forms, but simply focuses attention on surface exponence as a rich domain of generalization within morphology just as e.g., the phonological word is the domain of generalizations such as vowel harmony. In line with this Robbins 1959:127 observes that “the word as a unity is more easily susceptible to grammatical statements than is the individual bound form.” The hypothesis that periphrastic forms occupy paradigm cells is argued for in Ackerman and Stump 2004.

Type 1 (T1): ending in C (except a glottal) or V;

Type 2 (T2): subtype 1 (i): stem ends in nasalizing/voicing glottal stop (=h)

subtype 2: (ii) stem ends in non-nasalizing/devoicing glottal stop<sup>27</sup> (=q)

These Types are exemplified in the following tables, where shared stem shape is indicated by identical shading for cells:

	Singular	Plural	Dual
Nominative	<i>ngano</i>	<i>nganoq</i>	<i>nganoxoh</i>
Accusative	<i>nganomh</i>	<i>nganu</i>	<i>nganoxoh</i>
Genitive	<i>nganoh</i>	<i>nganuq</i>	<i>nganoxoh</i>
Dative-Directional	<i>nganonh</i>	<i>nganoxoq</i>	<i>nganoxoh nyah</i>
Locative-Instrumental	<i>nganoxona</i>	<i>nganoxoqna</i>	<i>nganoxoh nyana</i>
Ablative	<i>nganoxod</i>	<i>nganoxot</i>	<i>nganoxoh nyad</i>
Prolative	<i>nganowna</i>	<i>nganuqmana</i>	<i>nganoxoh nyamna</i>

Table 8: Type 1: polysyllabic vowel stem: *ngano* ‘boat’

	Singular	Plural	Dual
Nominative	<i>wih</i>	<i>wiq</i>	<i>wingh</i>
Accusative	<i>wimh</i>	<i>wingo</i>	<i>wingh</i>
Genitive	<i>wi<sup>h</sup></i>	<i>wingoq</i>	<i>wingh</i>
Dative-Directional	<i>windh</i>	<i>wingq</i>	<i>wingh nyah</i>
Locative-Instrumental	<i>wingana</i>	<i>wingaqna</i>	<i>wingh nyana</i>
Ablative	<i>wingad</i>	<i>wingat</i>	<i>wingh nyad</i>
Prolative	<i>wimna</i>	<i>wingoqmana</i>	<i>wingh nyamna</i>

Table 9: Type 2i: nasalizing glottal stem: *wih/wing* ‘tundra’<sup>28</sup>

<sup>27</sup>See Salminen 1997, 1998 for a careful taxonomy of stem types in Tundra Nenets which forms the basis for the analysis below. The orthographic conventions in the table represent an admixture of Latinized traditional Cyrillic orthography and Salminen’s phonological representation. This is intended to make the representations transparent without going into more phonological detail than the use of Salminen’s phonological representations would require. There are inevitably, as a consequence, certain aspects of the representations which are misleading. In contrast to the utilitarian motivations guiding the representations in these tables, all of the statistical calculations are based on Salminen’s phonological transcriptions of words.

	Singular	Plural	Dual
Nominative	<i>myaq</i>	<i>myadq</i>	<i>myakh</i>
Accusative	<i>myadmh</i>	<i>myado</i>	<i>myakh</i>
Genitive	<i>myadh</i>	<i>myadoq</i>	<i>myakh</i>
Dative-Directional	<i>myat</i>	<i>myakh</i>	<i>myakh nyah</i>
Locative-Instrumental	<i>myakana</i>	<i>myakaqna</i>	<i>myakh nyana</i>
Ablative	<i>myakad</i>	<i>myakat</i>	<i>myakh nyad</i>
Prolative	<i>myaqmna</i>	<i>myadoqmana</i>	<i>myakh nyamna</i>

Table 10: Type 2ii: nasalizing glottal stem: *myaq/myad* ‘hut’

Examination of these patterns yields a basic observation about Tundra Nenets: the relevant nominal paradigms for all stem classes are partitioned into subparadigms, each of which is defined by the presence of a characteristic and recurring stem (*ngano*, *nganu*, or *nganoxo*). In what follows we will refer to these forms as *recurrent partials* and the sets in which they recur as *coalitions* or *alliances* of forms. This permits one to offer the following generalization about Tundra Nenets absolute nominal paradigms:

Subparadigms are domains of interpredictability among alliances of wordforms, rather than of derivability from a single base wordform.<sup>29</sup>

In the next section we explore how this hypothesis based on recurrent parts fares against a competitor based upon derivation of forms from a single surface base form. It is important to keep in mind that both of these approaches share the intuition, in contrast to the syntagmatic/compositional appraised previously, that generalizations are built upon surface words. They differ, among other ways, with respect to the role that surface words play in identifying the correct generalizations.

#### 4 Competing surface-oriented hypotheses

An approach based on recurrent partials and patterns of relatedness among forms finds antecedence and inspiration in Bochner (1993), where it is argued that no form need necessarily serve as a privileged base form among different surface expression of single lexeme.

---

<sup>28</sup>The occurrence of a specific allomorph, e.g., *wingana*, where *-gana* is part of a family allomorphs such as *-xana*, *-kana*, leads to the inference that this word belongs to the class of stem final nasalizing glottals. Thus surface allomorphy can be used as a diagnostic clue for reliably guiding paradigm-based inferences.

<sup>29</sup> The need for access to inflected forms within (sub)paradigms for purposes of derivational relatedness is evident from the fact that at least two verbal derivation operations are built upon the form used to express genitive plural nominals. (See Kupryanova et al. 1985:139.)

Regardless of whether a stem exists as an independent word, all these systems share the property that they have clusters of related forms where it is at least somewhat arbitrary to take any one form as basic. This is what I take to be defining characteristic of a paradigm. Thus, we need a way to relate to the various members of paradigm directly to each other without singling out any one of them as a base for the others.” (1993:122)

On this type of analysis, alliances of wordforms share formal properties (i.e., recurrent partials), but the elements in such alliances need not be thought of as bearing derivational or constructive (in the sense of Blevins 2006) relations to one another, let alone to a single isolable base form. Derivational or constructive relations based on a single form can be described as *asymmetric*, since some specific form is assumed to be predictive of the other forms. In contrast, the present account can be described as a *symmetric* approach to the relations among members of (sub)paradigms, where there is no one form that serves as the base from which the others are derived (though the lack of a single privileged base does not entail that there cannot be multiple subparadigms in which a particular recurring form (a *partial*) serves a pivotal role). This basic organization is depicted in Figure 3, where the whole Tundra Nenets nominal declension paradigm is partitioned into three alliances of forms.

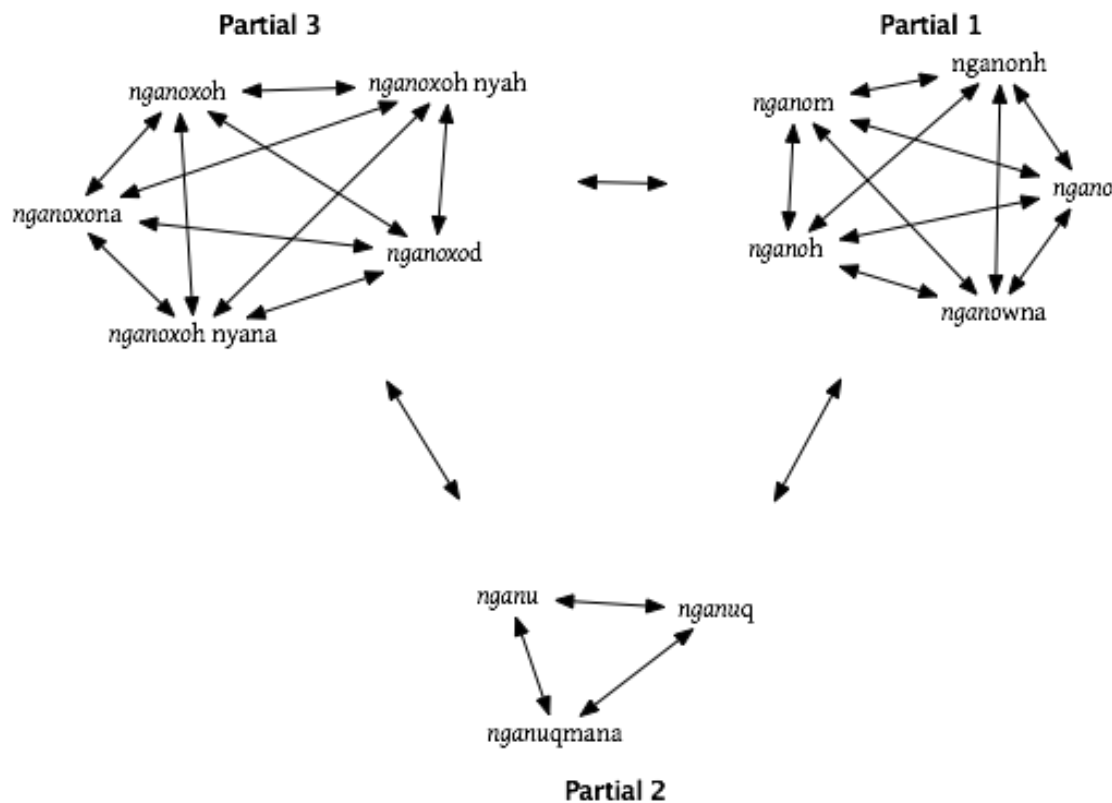


Figure 3: Symmetric paradigm organization

Each form in a subparadigm provides information about other forms in the same subparadigm. The members of a subparadigm share partials, thereby making the an alliance a system of interpredictability among related wordforms.

This view of the patterned nature of morphological organization as independent of claims concerning the need to identify a single base (either underlying or surface) from which rule applications construct related wordforms finds common ground with earlier approaches. Such an approach is raised by Anderson (1985:53), which identifies Jackendoff (1975) as ancestral to the sort alternative articulated here:

We must emphasize that while Saussure has no sympathy for a description of alternations which posited unitary underlying forms and rules altering the character of segments, he certainly considered alternations to be a rule governed aspect of sound structure... As such all of his rules have the character of lexical redundancy rules (in the sense of Jackendoff 1975)...

Indeed, Bochner (1993) explicitly develops the pattern-based insights offered in Jackendoff’s seminal article.

Turning to the more conventional and familiar strategy of an asymmetric approach, we will discuss Albright 2002 as a particularly cogent model. He writes (2002:118 )

“learners are restricted to selecting just one form as the base within the paradigm, and all other forms must be derived from the same base.”<sup>30</sup> (Albright 2002:118)

On this account, there is a single most informative surface form within any paradigm from which all other forms are derived. This single base form is the same for all conjugation or declension classes. Thus, if the partitive singular is identified as the base for declension class 1, then the partitive singular form must likewise serve as the base for all other declension classes.

The postulation of a base entails an asymmetric derivational relation between it and derived forms. This strong claim, however, is somewhat attenuated by the fact that Albright additionally suggests that the postulation of a single surface base may not preclude the possibility of multiple local bases: “When we look at larger paradigms... it often appears that we need local bases for each sub-paradigm (something like the traditional idea of *principal parts*, or multiple stems)” (Albright 2002:118).

The basic asymmetric organization with local bases may be depicted as in Figure 4:

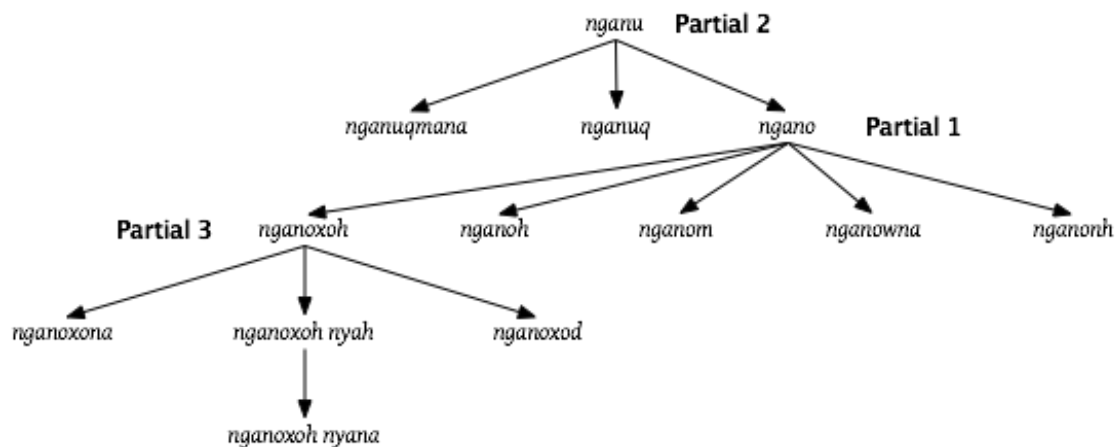


Figure 4: Assymmetric paradigm organization

Crucially, in contrast to what is shown in Figure 3, each subparadigm contains a base from which the rest of the forms in it are derived, and each local base is similarly asymmetric in the relations between wordforms. There is no notion of interpredictability of the sort manifest in Figure 3: the base gives information about derived form, but the derived form need not give information about base.

<sup>30</sup> This resembles one of the analytic options for determining (underlying) base forms argued against in Kenstowicz and Kissebirth 1979:201: “The alternant selected as the UR must occur in the same morphological category for all morphemes of a given morphological class (verb, noun, particle, etc.)”

As proposed, Albright’s model makes a strong cross-linguistic claim that language architecture exhibits asymmetric relations of the proposed types. Instead, we will take a less *aprioristic* stance and suggest that it is an empirical matter whether (portions of) a language’s morphological system is best analyzed as symmetrical or asymmetrical. The claim we defend in the remainder of this article will be that Tundra Nenets is symmetrical. Specifically, we hypothesize that Tundra Nenets nominal paradigms are organized into several subparadigms, and the domains in which recurrent forms bases play role constitute domains of interpredictability. The basic picture can be conceptualized as follows: subparadigms are systems of related wordforms organized around recurring partials. Partials need not be independently occurring words, though two of the three primary partials in Nenets happen to be free morphs. Within the subparadigm there is no reason to assume a derivational relation between the partials. These are not local bases (in Albright’s sense), but simply patterns (in Bochner’s sense).

Given these two competing and coherent alternatives, the question arises as to how they can be compared. The strategy that we have chosen is to explore whether the most challenging and problematic instance of relatedness between two wordforms is reliably asymmetric and based on the same morphosyntactic cell across declension classes, and so to test Albright’s Single Surface Base Hypothesis.<sup>31</sup> The logic is straightforward: if the complexity of analyzing the two least transparently related words within a paradigm can be simply stated by referring to a single constant base form across conjugation classes, then this will stand as an argument for the Single Base Hypothesis. Conversely, under the symmetric analysis, we expect that no single form can be compellingly demonstrated to be the base, but that a principled account can be stated in terms of patterns of interpredictability within alliances of forms.

## 5 A comparison of competing hypotheses

Consider the following pairs of NOM.SG and ACC.PL forms of related lexemes in Table 1.<sup>32</sup>

Nom sg.	Acc. pl.	Gloss
<i>ngøno</i>	<i>ngønu</i>	‘boat’
<i>lyabtu</i>	<i>lyabtu</i>	‘harnessed deer’
<i>ngum</i>	<i>nguwo</i>	‘grass’
<i>xa</i>	<i>xawo</i>	‘ear’
<i>nyum</i>	<i>nyubye</i>	‘name’
<i>yí</i>	<i>yíbye</i>	‘wit’
<i>myir</i>	<i>myirye</i>	‘ware’
<i>wíh</i>	<i>wíngo</i>	‘tundra’
<i>weh</i>	<i>weno</i>	‘dog’

<sup>31</sup> Stump & Finkel’s proposal concerning a dynamic strategy for implicative relations in paradigms suggests that the hypothesis of a single recurring and reliable cell across classes is incorrect.

<sup>32</sup> These wordforms are taken from Salminen 1997 and, consequently, presented in their original transcription where a superscripted <sup>o</sup> designates a schwa and  $\emptyset$  represents a reduced vowel. These comparisons are somewhat misleading, since they neither indicate syllabic cues nor type frequency associated with the pairs of wordforms. Efforts were made to control for these factors in the calculations described below.

<i>nguda</i>	<i>ngudyi</i>	‘hand’
<i>xoba</i>	<i>xob<sup>o</sup></i>	‘fur’
<i>saw<sup>o</sup>nye</i>	<i>saw<sup>o</sup>nyi</i>	‘magpie’
<i>tyírtya</i>	<i>tyírtya</i>	‘bird’

Table 1: Tundra Nenets inflected nominals

A comparison of the forms in the columns reveals that there is indeterminacy or uncertainty with respect to predictability in both directions. For example, while the ACC.PL of both *boat* and *harnessed deer* end in the vowel *-u*, their NOM.SG forms end in *-o* and *-u* respectively. Likewise, while the NOM.SG of *boat* ends in *-o*, it is the ACC.PL of *grass* ends in *-o*, while its NOM.SG ends in the consonant *-m*. Representative mappings for subset of the nominal inventory are provided in Figure 5.

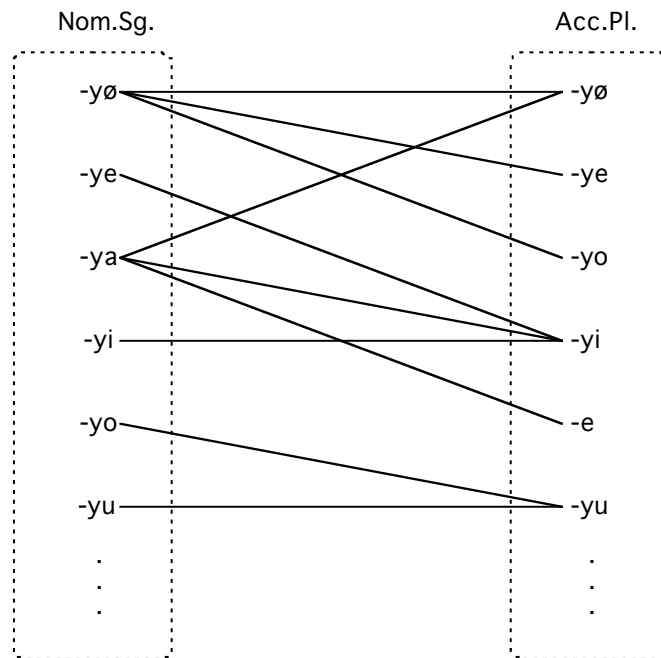


Figure 5: NOM. SG. and ACC. PL. pairings in Tundra Nenets

Given exposure to one form, how can one predict the other? This is the Paradigm Cell Filling problem. From the perspective of a single base hypothesis, the question is whether one is derivable from the other (or whether some 3<sup>rd</sup> form can serve as a base for both<sup>33</sup>). In the following pilot study we, explore the relative predictiveness of NOM.SG and ACC.PL, with the following basic question in mind: between the NOM.SG and the ACC.PL, which, if either, is more useful for predicting the other? The data for this analysis derives from the corpus of 4,334 nominals extracted from

<sup>33</sup> A comprehensive examination of Albright’s informativeness measure would test the entire Tundra Nenets paradigm in order to identify the most informative form. Given the maximal difference between these nom.sg. and acc.pl. forms, we simply consider directionality of derivation between them, ignoring other related wordforms. A more complete study would also explore whether deviations of predictability from a single base are sensitive to phonological cues that we many not have not coded for.

Salminen’s dictionary of 16,403 entries, based on Tereshchenko (1965). The dictionary specifies meaning, frequency, as well as the stem class assignment for lexemes.

In order to assess this relation, we will use the information theoretic notion *entropy* as the measure of predictability. This permits us to quantify “prediction” as a change in uncertainty, or information entropy (Shannon 1948). Suppose we are given a random variable  $X$  which can take on one of a set of alternative values  $x_1, x_2, \dots, x_n$  with probability  $P(x_1), P(x_2), \dots, P(x_n)$ . Then, the amount of uncertainty in  $X$ , or, alternatively, the degree of surprise we experience on learning the true value of  $X$ , is given by the entropy  $H(X)$ :

$$H(X) = -\sum_i P(x_i) \log_2 P(x_i)$$

The entropy  $H(X)$  is the weighted average of the *surprisal*  $-\log_2 P(x_i)$  for each possible outcome  $x_i$ . The surprisal is a measure of the amount of information expressed by a particular outcome, measured in *bits*, where 1 bit is the information in a choice between two equally probable outcomes. Outcomes which are less probable (and therefore less predictable) have higher surprisal. Specifically, surprisal is 0 bits for outcomes which always occur ( $P(x)=1$ ) and approaches  $\infty$  for very unlikely events (as  $P(x)$  approaches 0). The more choices there are in a given domain and the more evenly distributed the probability of each particular occurrence, the greater the uncertainty or surprise there is (on average) that a particular choice will be made among competitors and, hence, the greater the entropy. Conversely, choices with only a few possible outcomes or with one or two highly probable outcomes and lots of rare exceptions have a low entropy. For example, the entropy of a coin flip as resulting in either heads or tails is 1 bit; there is equal probability for an outcome of either heads and tails:

$$\begin{aligned} H(X) &= -\sum P(x_i) \log_2 P(x_i) \\ &= -(P(\text{heads}) \times \log_2 P(\text{heads}) + P(\text{tails}) \times \log_2 P(\text{tails})) \\ &= -(0.5 \times \log_2 0.5 + 0.5 \times \log_2 0.5) \\ &= 1 \end{aligned}$$

The entropy of a coin rigged to always come up heads, on the other hand, is 0 bits: there is no uncertainty in the outcome:

$$\begin{aligned} H(X) &= -\sum P(x_i) \log_2 P(x_i) \\ &= -(P(\text{heads}) \times \log_2 P(\text{heads}) + P(\text{tails}) \times \log_2 P(\text{tails})) \\ &= -(1.0 \times \log_2 1.0 + 0.0 \times \log_2 0.0) \\ &= 0 \end{aligned}$$

With this as background we can now measure the entropy among Tundra Nenets nominal types. For present purposes, we identify 31 different types of nominative singular nouns. Some, like the class of words ending in the reduced vowel  $-\emptyset$ , are quite common, while others are quite rare. Overall, the entropy of this distribution  $H(\text{NOM.SG})$  is 3.28 bits. This means that it requires this many bits to account for all of the nominative norms in terms of the likelihood of particular

form being encountered. Similarly, there are 35 different types of accusative plurals, and their entropy  $H(\text{ACC.PL})$  is 3.36 bits.

Having quantified the degree of uncertainty in the choice of NOM.SG and ACC.PL types individually, we can now calculate predictability of one type given the other: we can measure the size of the surprise using *conditional entropy*  $H(Y|X)$ , the uncertainty in the value of  $Y$  given that we already know the value of  $X$ . The smaller  $H(Y|X)$  is, the more predictable  $Y$  is on the basis of  $X$ , i.e., the less surprised one is that  $Y$  is selected. In the case where  $X$  completely determines  $Y$ , the conditional entropy  $H(Y|X)$  is 0 bits: given the value of  $X$ , there is no question remaining as to what the value of  $Y$  is. On the other hand, if  $X$  gives us no information about  $Y$  at all, the conditional entropy  $H(Y|X)$  is equal to  $H(Y)$ : given the value of  $X$ , we are just as uncertain about the value of  $Y$  as we would be without knowing  $X$ .

Given our competing hypotheses, we would expect the following. Under the asymmetry hypothesis, we should see a greater degree of predictability (and a lower conditional entropy) in one direction than in the other, meaning that either NOM.SG. or ACC.PL should be a good predictor of the other.<sup>34</sup> In contrast, the absence of a global directionality of predictability is compatible with the symmetric approach, though additional evidence would be required to address its claims of alliances of interpredictability. Thus, the present investigation focuses on the viability of grammatical architectures that assume asymmetry and a single base.

To test these alternatives we make the following comparisons. First, consider predicting the ACC.PL form from the NOM.PL of a given word. We can evaluate the difficulty of this prediction using the conditional entropy  $H(\text{ACC.PL}|\text{NOM.SG})$ , the uncertainty in the ACC.PL given the NOM.SG. Out of the  $31 \times 35 = 1,085$  possible pairings of NOM.SG and ACC.PL types, 52 are actually attested in the lexicon. In some cases, knowing the NOM.SG of a word uniquely identifies its ACC.PL, e.g. a word ending in  $-ye$  in the NOM.SG always has an ACC.PL in  $-yi$ . For such words, once we know the NOM.SG there is no uncertainty in the ACC.PL and the conditional entropy  $H(\text{ACC.PL}|-ye) = 0$  bits. In other cases, however, knowing the NOM.SG narrows down the choices for the ACC.PL but does not uniquely identify it. For example, words whose NOM.SG ends in  $-ya$  might have an accusative plural in  $-\emptyset$ ,  $-yi$ ,  $-y\emptyset$ , or  $-e$ . On average, across the whole (sample) lexicon, the uncertainty in the ACC.PL given the NOM.SG is 0.59 bits. In other words, the NOM.SG “predicts” all but 0.59 of the 3.36 bits of uncertainty previously calculated for the ACC.PL. Now, if we switch directions, going from ACC.PL to NOM.SG, it turns out that the conditional entropy  $H(\text{NOM.SG}|\text{ACC.PL}) = 0.51$ . In other words, the ACC.PL “predicts” all but 0.51 of the 3.28 bits in the NOM.SG. Since the conditional entropy is closer to 0 in the latter than in the former, the ACC.PL appears to be slightly more helpful for predicting the NOM.SG than vice versa, but only by a slim margin. More importantly, neither conditional entropy is 0 bits or close to it, meaning neither for is especially useful for predicting the other. Thus, we can conclude, at least provisionally, that the postulation of a single

---

<sup>34</sup> Of course, if it turns out that only some forms reliably predict other form unidirectionally, i.e., there is low conditional probability going in one direction for select number of instances, but not for others, one might argue that the offending pairs are lexically listed exceptions and, consequently, uninteresting. We believe that this strategy runs the risk of ignoring small regularities by overwhelming them with possibly less representative more transparent pairs of relations.

base is not warranted for these data and that, accordingly, the asymmetric proposal seems overly restrictive as a universal principle of paradigm organization.

The preceding calculations were based on *type* frequencies in which each distinct noun lexeme was counted once. We can also calculate conditional entropy on the basis of *token* frequencies, which consider all occurrences of nouns in the corpus, where a given type can account for a large number of tokens. Using the token counts from Salminen's corpus, we arrive at the following entropy values:

$$\begin{aligned} H(\text{NOM.SG}) &= 3.81 \text{ bits} \\ H(\text{ACC.PL}) &= 4.02 \text{ bits} \\ H(\text{ACC.PL}|\text{NOM.SG}) &= 0.91 \text{ bits} \\ H(\text{NOM.SG}|\text{ACC.PL}) &= 0.70 \text{ bits} \end{aligned}$$

As is evident, attention to the token frequencies yield essentially the same conclusions as those derived from consideration of type frequencies: there is significant surprise associated with prediction in either direction, and neither form predicts the other very well. Thus, whether one counts types or tokens, hypothesizing one form or the other as the single privileged base would be arbitrary and would entail a large inventory of irregular residual pairings to be memorized by the language learner. This arbitrariness is avoided on a symmetric account, where there is no need to suppose that some forms are reliably predictable from another form in a single direction. Rather, as presented previously, the symmetrical proposal posits alliances which cohere into interpredictable coalitions of forms and which together partition the entire paradigm. We do not expect forms which take part in different sets of alliances to be mutually predictive on this account, so the results from conditional entropy concerning the absence of reliable directionality is not surprising.

But, do the Tundra Nenets distributions of form bear on the proposal of alliances? Acknowledging the essential arbitrariness of positing either form as the base, it still turns out that one is far more likely to encounter a NOM.SG form of a given word than the ACC.PL form. This is evident from the frequency distributions of the absolute declension for all case and number encodings of the 12,152 noun tokens in Salminen's sample sentence corpus:<sup>35</sup>

---

35 This corpus contains 9,993 sentences consisting of 39,417 words. Note that this corpus consists of example sentences from the Nenets/Russian dictionary rather than cohesive narrative texts or discourse, so the frequencies reported here may not be completely representative of the natural speech that serves as input for learning.

	sg	du	pl
nom	4,117	7	770
gen	3,002	6	376
acc	1,077	5	355
dat	762	0	89
loc	724	0	108
abl	291	0	50
pros	372	0	41

Table 12: Wordform frequencies in Tundra Nenets

In Table 12, the NOM.SG represents 33.8% of the tokens, while the ACC.PL represents only 2.7%. If the most frequent form was the most useful for solving the Paradigm Cell Filling problem speakers assume the NOM.SG is the base, but recall that this does not correspond to the predictiveness of forms as calculated previously; neither NOM.SG nor ACC.PL is a reliable predictor of the other across the whole nominal inventory. Correlatively, if it happened that the results from conditional probability calculations suggested that ACC.PL was a likely base, or if we arbitrarily posited this form as the unidirectional base, the attested frequencies suggest that for any given lexeme speakers would have a low probability of actually encountering the purportedly diagnostic ACC.PL form. The situation is even worse for forms such as the direct case dual forms, which account for 0.1% of the tokens. In fact, no individual wordform (except for the NOM.SG and the GEN.SG) occurs with high enough frequency to be a reliable source of information about a word's inflectional class. From a lexicographic point of view, this makes Tundra Nenets a very challenging language, as one might never encounter the diagnostic forms necessary to identify a word's inflectional class. However, the issue takes on a different shape when we look at the forms in terms of alliances. In this connection, it is crucial to observe that the form with the second highest frequency GEN.PL is transparently related to ACC.PL forms: the former is always identical to the latter with the addition

of final glottal stop.<sup>36</sup> Thus, *ngønu* is ACC.PL form for *boat*, while *ngønuq* is its GEN.PL. From this perspective, the NOM.SG belongs to Partial 1, ACC.PL to Partial 2, and e.g., the direct dual forms belong to PARTIAL 3. While it is clear that there is a low likelihood of encountering ACC.PL, there is a much higher likelihood of encountering the partial associated with ACC.PL if we posit subparadigms of unpredictable forms, particularly given the relatively high frequency of GEN.PL forms located in this alliance. Each of the wordforms based on Partial 2 provides information about the shape of all of the others. If we sum the token frequency distributions for all absolute and possessive forms in terms of partials, we get the following:

Partial 1	13,083
Partial 2	1,717
Partial 3	1,782

Thus, speakers can gain fairly reliable cues about the shape of even very low frequency wordforms if they rely on alliances of related forms within subparadigms. It is important to note at this juncture that even if a derivability relation had been assumed, counter to available evidence, this would not have accounted for the evident subpatterns of shared forms as described above, rendering such subpatterns epiphenomenal, rather than central to organization and the solution of the cell filling problem as on the present account.<sup>37</sup>

In sum, both type and token calculations suggest that, neither the NOM.SG nor ACC.PL form reliably serves as the single base from which the other is predicted. These equivocal results with respect to directionality of prediction contrast with the overwhelming asymmetry of the probability of encountering NOM.SG versus ACC.PL on the basis of frequency. Positing subparadigms reveals that partials appear with much higher frequency than any given wordform, so that there is no need to encounter a specific form in order to predict allied forms. What is important, by hypothesis, is that the aggregate frequency of partials be high enough to be useful. And if so, then the paradigm cell filling problem is solved.

## 6 Provisional conclusions and ramifications

We have seen that Bochner's symmetrical pattern sets and Albright's asymmetric local bases can both be used to model paradigm structure, but the two models make very different predictions when considered in the light of the Paradigm Cell Filling problem. Under Albright's model, derived forms should be predictable from the base form, but there is no reason to expect bases to be pre-

---

<sup>36</sup> Since the ACC PL and GEN PL forms are transparently related, the entropy calculations outlined above apply to both forms equally. So, even though the GEN PL is more frequent than the ACC PL, it is still not a reliable predictor of the NOM SG form.

<sup>37</sup> Though we have no explanation for why the alliances consist of the particular forms and feature sets they do, it is also not necessary clear that this matters for synchronic purposes of filling paradigm cells.

dictable from derived forms or derived forms to be predictable from each other. Under Bochner's model, on the other hand, we expect potentially complex interrelations among forms in the same paradigm or subparadigm. (Sub)paradigms are organized in terms of patterns of whole wordform relatedness with members of (sub)paradigms exhibiting interpredictability: this facilitates solving the Paradigm Cell Filling problem. Despite the complexity of the paradigm, one can reliably predict an inflected form of a word given any exemplars other inflected forms from alliances in languages like Tundra Nenets. Note that we have focused here, in effect, on the role that analogy plays in organizing wordforms in synchronic paradigms and how this may facilitate learning, and not on the role that analogy may play in driving changes within (sub)paradigms. Albright's model seeks confirmation mainly in these diachronic effects, while we presently have no speculations about this. On the other hand, as observed by Albright (2007), a model of the present sort makes a robust synchronic prediction which distinguishes it from his own model: if alliances or coalitions of interpredictability are real, there should be psycholinguistic evidence that reveals a reliance on them. In fact, evidence of this sort has been found by Millin et al. (2007), in which recognition times for one wordform are affected by the relative frequencies of other wordforms in that lexeme's paradigm.

Finally, we have focused here on an essentially developmental issue concerning the induction of generalizations, and have developed a pattern-based learning strategy which operates on networks of whole words. This is consistent with developmental research that suggests a trajectory involving increasing levels of schematicity from the concreteness of individual instances to general patterns capable of relating them to other instances across various dimensions: children are exposed to specific instances and then discover/develop the complex configurations associated with particular words as well as the relatedness schemata of increasing abstractness which license inferences about novel wordforms (Tomasello 2003; Gentner and Namy 2004; Pinker 1984; MacWhinney 1978, among others).<sup>38</sup> Of course, this leads to questions concerning how children go about identifying the relevant dimensions of morphosyntactic properties that receive surface expression in words and how they isolate the appropriate patterns of surface exponence. Crucially, it becomes clear that hypotheses concerning the Paradigm Cell Filling problem have consequences for how we formulate acquisition questions, among other sorts of external evidence bearing on linguistic analysis. Just as the syntagmatic/compositional approach to morphology has entailed certain research questions and encourages certain types of explanations within linguistics and related disciplines, the paradigmatic/configurational approach yields a different set of questions and explanations, many clearly coincident with the recent research directions and results in other disciplines within recent cognitive(neuro)science. It suggests that we identify the assumptions and methodologies that have led to reliable results and well-founded theories in other disciplines and inquire as how they may inform linguistic analysis. For a long time linguistics has focused on how language is different from other phenomena, even radically so. It seems time to inquire as to how language can be analyzed as elementally similar to many other complex phenomena, without fear of losing its uniqueness.

---

<sup>38</sup> This also leads to further research exploring influences on learnability via connectionist modeling of the paradigm cell filling task. (Thymé 1993, Thymé, Ackerman, and Elman 1994, Goldsmith and O'Brien 2006)

## References

- Ackerman, F. and G. Stump. 2003. Paradigms and periphrasis: A study in realization-based lexicalism, In A. Spencer and L. Sadler eds. *Projecting morphology*. CSLI Publications.
- Ackerman, F. and G. Webelhuth. 1998. *A Theory of Predicates*. CSLI Publications.
- Albright, A. 2002. The identification of bases in morphological paradigms. UCLA doctoral dissertation.
- Albright, A. & B. Hayes. 2002. Islands of reliability for regular morphology: Evidence from Italian. *Language* 78: 684-709.
- Albright, A. & B. Hayes. 2003. Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition* 90: 119-161.
- Anderson, S. R. 1985. *Phonology in the 20<sup>th</sup> century*. University of Chicago Press.
- Anderson, S.R. 1992. *A-morphous morphology*. Cambridge University Press.
- Anderson, S. R. & D. W. Lightfoot. Biology and Language: A response to Everett (2005). *Journal of Linguistics* 42(2):377-383.
- Aronoff, M. 1994. *Morphology by itself*. MIT Press.
- Berent, I., & Pinker, S. (in press). [The Dislike of Regular Plurals in Compounds: Phonological or Morphological](#). *The Mental Lexicon*.
- Blevins, J. 2006. Word-based morphology. *Journal of Linguistics* 42(3).
- Blevins, J. 2005. Word-based declensions in Estonian. *Yearbook of Morphology 2005*, G. Booij & J. van Marle (eds.), Dordrecht: Springer, 1–25.
- Blevins, Juliette. 2004. *Evolutionary Phonology: The emergence of sound patterns*. Cambridge University Press.
- Bochner, H. 1993. *Simplicity in generative grammar*. Mouton de Gruyter.
- Booij, G. 2005. *The grammar of words: An introduction to linguistic morphology*. Oxford University Press.
- Bybee, J.L. 1985. *Morphology: a study of the relation between meaning and form*. Johns Benjamins.
- Camazine, S, J.-L. Deneubourg, N. R. Franks, J. Sneyd, G. Theraulaz, and E. Bonabeau. 2001. *Self organization in biological systems*. Princeton University Press,
- Carroll, S. 2005. *Endless forms most beautiful*. Norton Publications.
- Cech, P. 1995/1996. Inflection/derivation in Sepečides-Romani. *Acta Linguistica Hungarica* 43.1–2:67–92
- Clahsen, H. 1999. Lexical entries and rules of language: A multidisciplinary study of German inflection. *Behavioral and Brain Sciences* 22: 991–1060.

- Clahsen, H. and M. Almazan. 2001. Compounding and inflection in language impairment: Evidence from Williams Syndrome (and SLI). *Lingua* 111:729–757.
- Clahsen, H., M. Ring, and C. Temple. 2003. Lexical and morphological skills in English-speaking children with Williams Syndrome. Ms.
- Corning, P. 2005.
- Durand, J. & B. Laks. 2002. Phonology, phonetics, and cognition. In J. Durand & B. Laks eds. *Phonetics, Phonology, and Cognition (Oxford Studies in Theoretical Linguistics)*. Oxford University Press.
- Embick, D. 2007. Blocking effects and analytic/synthetic alternations. *Natural Language and Linguistic Theory* 25.1:1-37.
- Embick, D. & R. Noyer. 2005. Distributed morphology and the morphology/syntax interface. In G. Ramchand and C. Reiss eds., *The Oxford Handbook of Linguistic Interfaces*, Oxford University Press
- Gentner, D., & L. L. Namy. 2004. [The role of comparison in children's early word learning](#). In D. G. Hall & S. R. Waxman (Eds.), *Weaving a lexicon* (pp. 533-568). Cambridge: MIT Press.
- Gurevitch, O. 2006. Constructional morphology: The Georgian version. UC Berkeley PhD Dissertation.
- Goldsmith, J. and J. O'Brien. 2006. Learning inflectional classes. To appear in *Language Learning and Development* 24(4): 219-250.
- Jablonka, E. and M. J. Lamb. 2005. *Evolution in Four Dimensions: Genetic, Epigenetic, Behavioral, and Symbolic Variation in the History of Life*. MIT Press.
- Keller, Evelyn Fox. 2002. *Making Sense of Life*. Harvard University Press.
- Keresztes, L. 1990. *Chestomatica Morduinica*. Budapest: Tanykönyvkiadó.
- Keresztes, L. 1999. *Development of Mordvin Definite Conjugation*. Suomalais-Ugrilainen Seura: Helsinki.
- Kirby, J. 2006. Minimal redundancy in word-based morphology. University of Chicago Ms.
- Kupryanova, Z. N. et. al. 1985. *Nenetskij jazyk [Nenets Language]*. Nauk: Moscow.
- MacWhinney, B. 1978. *Processing a first language: the acquisition of morphophonology*. Monographs of the Society for Research in Development 43.
- Matthews, P. H. 1991. *Morphology*. Cambridge University Press.
- Mészáros, Édit. 1998. *Erza-mordvin nyelvkönyv [Erza-Mordvin Pedagogical Grammar for beginners and intermediates]*. Szeged, Hungary: JATE Press.
- Neuvel, S. and S. Fulop. 2002. Unsupervised learning of morphology without morphemes. *Proceedings of the ACL-02 workshop on Morphological and phonological learning*. Pp. 31–40.
- Oudeyer, P-Y 2006. *Self-organization in the Evolution of Speech*. Studies in the Evolution of Language. Oxford University Press.
- Paunonen, Heikki 1976. Allomorfien dynamiikkaa [The dynamics of allomorphs]. *Virittäjä* 79:82–107.

- Pertsova, K. 2004. Distribution of genitive plural allomorphs in the Russian lexicon and in the internal grammar of native speakers. Master's thesis, UCLA, Los Angeles.
- Pennington, B. F. 2001. Genes and brain: Individual differences and human universals. In M. Johnson, Y. Munakata, Rick. O. Gilmore eds., *Brain development and cognition*. Blackwell Publishers.
- Pinker, S. 1984. *Language learnability and language development*. Harvard University Press.
- Pinker, S. 2000. *Words and rules: The ingredients of language*. Harper Perennial.
- Scholz, B. and G. K. Pullum. 2005. Irrational nativist exuberance. Ms.
- Ramscar, M. 2005. Learning language from the input: Why innate constraints aren't needed in compounding. Ms, Stanford University.
- Robins, R. H. 1959. 'In defence of WP', *Transactions of the Philological Society*, 116—44.
- Salminen, T. 1997. *Tundra Nenets inflection*. Helsinki : Suomalais-Ugrilainen Seura.
- Salminen, T. 1998. *A morphological dictionary of Tundra Nenets*. Helsinki: Suomalais-Ugrilainen Seura.
- Saussure, de., F. 1966. *Course in General Linguistics*. McGraw-Hill Publishers.
- Sapir. E. 1921. *Language: An Introduction of the Study of Speech*.
- Scholz, B. C. & G.K. Pullum. 2006. Irrational nativist exuberance. In R. Stainton ed. *Contemporary debates in cognitive science*. Basil Blackwell. 59-80.
- Shannon, C. E. 1948. A mathematical theory of communication. *Bell System Technical Journal*, vol. 27, pp. 379-423 and 623-656, July and October.
- Skousen, R. 1989. *Analogical modeling of language*. Dordrecht: Kluwer.
- Smith, L. & E. Thelen. 2003. Development as a dynamic system. *Trends in Cognitive Science*. Vol 7. No. 8:343-348.
- Spencer A. and L. Sadler. 2000. [Syntax as an exponent of morphological features](#). *Yearbook of Morphology 2000*.
- Stump, G. 2001. *Inflectional Morphology*. Cambridge University Press.
- Stump, G. 2002. Morphological and syntactic paradigms: Arguments for a theory of paradigm linkage. *Yearbook of Morphology 2001*.
- Stump, G. & Finkal. Principal parts and degrees of paradigmatic transparency. Ms.
- Tereshchenko, N. M. 1965/2003 *Nenetsko-ruskii slovar' [Nenets-Russian dictionary]*. St. Petersburg.
- Trommer, J. 2003. Hungarian has no portmanteau agreement. Ms.
- Thymé, A. 1993. Connectionist approach to nominal inflection: Paradigm patterning and analogy in Finnish. UC San Diego doctoral dissertation..
- Thymé, A. , F. Ackerman & J. Elman. 1994. Finnish nominal inflection: paradigmatic patterns and token analogy. *The Reality of Linguistic Rules*, Amsterdam: John Benjamins.
- Tomasello, M. 2003. *Constructing a Language: A Usage-based Theory of Language Acquisition*. Harvard University Press

- Trosterud, T. 2004/To appear. *Homonymy in the Uralic argument agreement systems*. Suomalais-Ugrilainen Seura, Helsinki. Finland.
- Väntillä, S. and F. Ackerman. 2000. Compounding and inflection in Finnish child language. *The Proceedings of the 30th Annual Child Language Research Forum*. Cambridge University Press.
- West-Eberhard, 2003. *Developmental Plasticity and Evolution*. Oxford University Press.
- Whorf, B. L. 1956. Science and linguistics. In John. B. Carroll ed. *Language, Thought, and Reality*. MIT Press.