Patterns of relatedness in complex morphological systems and why they matter

Farrell Ackerman
University of California at San Diego
fackerman@ucsd.edu

Robert Malouf
San Diego State University
rmalouf@mail.sdsu.edu

Workshop on Analogy in Grammar: Form and Acquisition
Department of Linguistics, Max Planck Institute for Evolutionary Anthropology
Leipzig, Germany
22–23 September 2006

> "One of the mysteries of biology is how the enormous amount of morphogenic, physiological, and behavioral complexity of an organism can be achieved with the limited amount of genetic information available within its genes. ... One of our fascinations with self-organization is its ability to create complexity from simplicity with remarkable economy."
>
> (Camazine, *et. al.* 2001:491)[1]

I. Our modest goal

1. Identify an important and largely neglected research area within morphological theory and morphological learning:

   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. I.e., given a novel inflected word form, what are all the other wordforms in its inflectional (and derivational) families?

2. Languages ordinarily depart from simple content/form mappings, some quite dramatically, and raise some questions:
   a. How are such complex systems organized?
   b. What role does this organization play with respect to licensing inferences concerning paradigm cell filling?
   c. What relation does this organization and the possibility for inferences have concerning the learnability of such complex systems?
3. Illustrate how Uralic (Finnish, Mordvin, Estonian, and mostly Tundra Nenets) provides fertile ground for identifying the nature of the challenges posed by this problem and clues for pos-

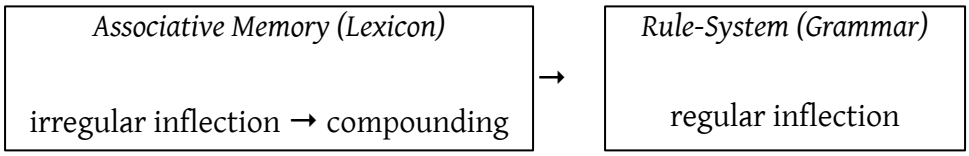
1 The references to biology are intended to contextualize morphological analysis within the "EvoDevo" spirit of J. Blevins' (2004) explorations in phonology.

sible solutions. We offer a provisional set of results about paradigm organization in Tundra Nenets.[2]

II. The landscape

A cautionary tale to motivate the need to look at complex morphological systems from a systems perspective

4. Of *mice-killers* and **rats-killers*[3]*:*
    a. Purported categoricality 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: Pinker (2000), Clahsen and Almazan (2001), Clahsen, *et al.* (2003).
    b. *Dual Route Model of Morphology*: No regular infection internal to derivation



    c. The plural of *mouse* is the irregular form *mice* and, hence, stored in the lexicon: it cannot be created by the application of a regular (=default) rule. Since it is stored, the compounding process in the lexicon has access to it. 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 ensuring that *rat-killer* can be formed. On the other hand, the regular plural formation operation can apply to compounds and yield, if need be, *rat-killers*
5. This (putative) split, and the earliness of behavioral (acquisition) evidence for it, suggest a biological constrains on grammar architecture with consequences for synchronic grammar organization and language learning: "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" (Clahsen 1999:1009)[4]
6. But, there *is* regular inflection internal to derivation in synchronic systems: In Sepečides-Romani (Cech 1995/1996:78), plural nominal forms can serve as bases for denominal verb derivation:
    a. Template for inchoative verb formation:

$$\text{NOUN}_{\text{plural}} + \textit{ndivola} \rightarrow \text{VERB}_{\text{get full of noun}}$$

---

2 The fieldwork on Tundra Nenets is facilitated by a Hans Rausing Endangered Language Major Documentation Project Grant 2003–2006 for which Ackerman is extremely grateful.

[3] See Ramscar (2005) on the (un)reliability of the judgments producing these data – they are claimed to be elicitation procedure artifacts.

[4] See Scholz and Pullum (2005) for a carefully reasoned critical evaluation of reflexive appeals to innateness of the sort guiding Clahsen's remarks.

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

Singular	Plural		
rukh 'tree'	*rukha* 'trees'	*O veš*	*rukhandivola*
		the wood	tree.PL.INCHOATIVE
		'The wood gets dense with trees.'	
džuv 'louse'	*džuva* 'lice'	*O bala*	*džuvandivola*
		the hair	louse.PL.INCHOATIVE
		'The hair is getting full of lice.'	

c. Formations apply to loanwords, depending only on whether the semantics of the prospective predicate is acceptable (loanword from Turkish): (Cech 1995/1996:78)
kurti 'worm' *kurtja* 'worms' *kurtjandivola* 'become full of worms'

d. 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).[56]

e. Cross-linguistic morphological research reveals that synchronic language systems do exhibit regular inflection internal to derivation, so they had better be learnable!

7. And, there *is* regular inflection internal to derivation in language acquisition

a. Results from diary study and experiments in Finnish language morphological learning (Oulu dialect): (Väntillä and Ackerman 2000)

b. Finnish nominal inflection: There are 15 case suffixes, two numbers (SG & PL), compounds are right-headed (N N_H), and non-heads tend to be uninflected (= NOMINATIVE[7]), GENITIVE-SG or GENITIVE-PL, but also can occur in various OBLIQUE cases.

c. Synchronic targets in Finnish:

SG.NOM	*käsi-∅-kauppa*	'over-the-counter sales'
SG.GEN	*käde-n-puristus*	'handshake'
PL.GEN	*käs-ien-hoitaja*	'manicurist',
SG.ELAT	*käde-stä-ennustaja*	'hand-reader'
PL.ADESS	*käs-illä-seisonta*	'handstand'

d. Finnish children use regular (=default) case inflection internal to lexical compounds. So, children had better not be broadly prohibited from acquiring regular inflection internal to derivation!

e. Indeed, Väntillä and Ackerman speculate that the existence of systems with greater complexity, may facilitate the learning of such systems by priming children to be sensitive to more distinctions and patterns earlier.

f. So, learning about complex systems is a necessary aspect for identifying generalizations likely to be true.

---

[5] Cech (1995/1996:80) observes that inchoative derivation using collectives such as 'smoke', suffixes the verb forming marker to the basic stem of collectives, since collective nominals don't have plurals. This clearly indicates that plural marking is not merely a formal marker used to form inchoatives.

[6] The genitive plural is used in various denominal verbal derivations in Tundra Nenets.

[7] Uninflected or NOMINATIVE marked non-heads represent approximately 75% of established lexical items. As seen below, however, inflected forms are quite productive.

8. Anyway, is it even intelligible to posit a gross constraint against inflection internal to derivation? All inflection is not the same, and some seems more derivational than others. It appears there is a gradient between inflectional and derivational morphology: (Booj 2002, 2005, Bybee 1985 among others)

$$\text{Root} - \text{derivation} - \text{inflection}_{\text{inherent}}[8] - \text{inflection}_{\text{contextual}}$$

9. The generative default: *Monogenetic theory of inflection*
    a. The monogenetic theory of inflection, like biological proposals for monogenic disorders, posits a determinate relation between gene(s) and some outcome, here, inflection in grammar.
    b. "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." (Jablonka & Lamb 2005: 57)
10. The speculative alternative: *Epigenetic theory of inflection*
    a. "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)

III. The basic analytic problem

Paradigm Cell Filling as prediction

11. 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" (Trosterud 2004:54)
12. What's a paradigm?
    a. "The PARADIGM of lexeme L is the set of pairs of morphosyntactic [grammatical] words and wordforms that realize L." (Trosterud 2004:15)

[8] Inherent refers to morphosyntactic categories such as semantic case and number, while contextual refers to morphosyntactic categories such as agreement and concord.

b. A (multidimensional) matrix of morphosyntactic properties whose cells define legal combinations of features for lexemes of specified categoriality (= *morphosyntactic/ grammatical word*) and whose occupants are surface wordforms of lexemes.[9]

13. Mappings between content and form
    a. Sometimes we find a one-to-one mapping between content and form ('agglutinating' languages)
    b. But, we also find more complex many-to-many relations between morphosyntactic property sets and wordforms[10], with the same formal pieces used for different functions (homonymy/syncretism)
    c. For example, in Tundra Nenets, the same paradigm (i.e., members of a suffix set) can be used with different lexical categories, sometimes serving essentially the same function, and sometimes serving different functions:[11][12]

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

14. Consequence: Words are *recombinant gestalts*, not simple (or even complex) combinations of bi-unique content-form mappings (i.e., morphemes).

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

Exponence of Tundra Nenets verbal forms as a function of conjugation

[9] Matthews 1991; Stump 2001, 2002; Ackerman & Stump 2003, Trosterud 2004, among others.

[10] 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.

[11] This discussion follows the presentation in Salminen (1997:.96; 103; 126).

[12] 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.

Examples:

General finite stem:

(i) subjective:	*tontaø-d^{0}m*
	cover.I (=1sg.)
	'I cover (something)'
(ii) objective sg.:	*tontaø-w^{0}*
	cover.II (=1sg/sg)
	'I cover it'

Dual obj. Stem:

(iii) objective du.	*tonta-gax^{0}yu-n^{0}*
	cover.dual.III (=1sg/du.)
	'I cover them (two).'

Special finite stem:

(iv) objective pl.	*tonteyø-n^{0}*
	cover.III (=1sg/pl.)
	'I cover them (plural)'
(v) reflexive	*tonteyø-w^{0}q*
	cover.IV (=1sg)
	'I got covered'

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

15. Given this "recombinant" quality of morphology, what can we say about natural language morphological systems and "optimal" organization? Before we say much we should better understand extraordinary complexifications of morphological systems such as[13]
    a. the wild profusion of nominal declension classes in Estonian relative to earlier Finnic and Uralic nominal declension (see Blevins 2005)
    b. the extraordinary and unprecedented articulation of agreement dimensions and attendant surface exponence in Mordvin verbal conjugation relative to all other Uralic systems.
16. Uralic object agreement correspondences between person/number of subj and person/number of obj (Keresztes 1999)
    a. Hungarian object agreement: Isg/2; 1 & 2 & 3/3
    b. Vogul, Osytak, Tundra Nenets object agreement: 1 & 2 & 3/3sg & 3 du & 3 pl

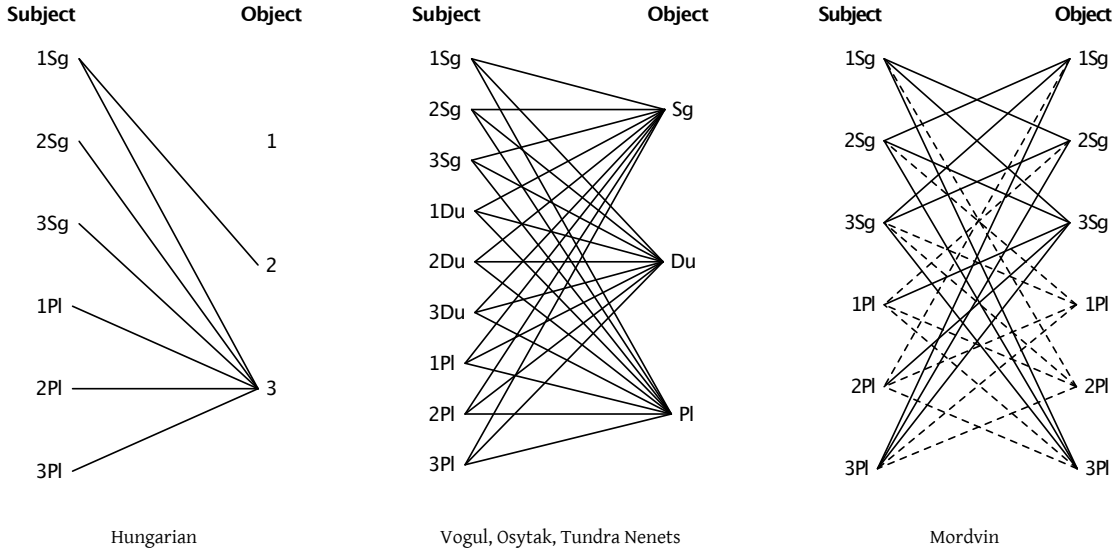
[13] We mention only Uralic here, but the reader should also consult Gurevitch (2006) for a careful examination of Georgian in this connection.

c. Mordvin object agreement (Mészáros 1998)
Erza Mordvin: *palams* 'to love' (following Keresztes 1990:46)

Subj/Obj → ↓	3sg	3pl
Sg 1	*pala-sa* 'I love him/her'	*pala-siń*
2	*pala-sak* 'you love him/her'	*pala-siť*
3	*pala-si*	*pala-sińźe*
Pl 1	*pala-sińek*	*pala-sińek*
2	*pala-sink*	*pala-sink*
3	*pala-siź*	*pala-siź*

Subj/Obj → ↓	1sg	1pl
Sg 2	*pala-samak* 'you love me'	*pala-samiź*
3	*pala-samam* 'he loves me'...	*pala-samiź*
Pl 2	pala-samiź	*pala-samiź*
3	pala-samiź	*pala-samiź*

d. A comparison (adapted from Keresztes 1999):



e. It is worth noting that there is systemic cost associated with the Mordvin innovation. The objective agreement paradigms in Hungarian, Vogul, Ostyak, and Nenets are all identical to some case/number paradigm for possessives in those languages (e.g., set

II above), whereas we do not find this correspondence between the objective agreement paradigm and the possessive paradigm in Mordvin.

f. Is there any presently useful sense in which these systems are more or less optimal than their historically antecedent systems (which may or may not have had object agreement at all) or the synchronic systems of related languages?

17. So, this is what it means to be complex wordforms in complex networks of relations.
    a. What are the principles of organization for these complex systems and what are the mechanisms of learning which permit them to be learned?
    b. There is too much specificity for much to be contentfully posited as "innate."
18. Paradigm Cell Filling: A (fairly) simple example
    a. For languages with several distinct inflectional classes, identifying which class a word belongs to is enough to complete its paradigm: Principal parts (Stump and Finkel)
    b. Finnish (Uralic) lexemes of category N(oun) are associated with the feature sets for case {nom, gen, part...}. num {sg,pl}....[14]
    c. Schematic partial paradigm for Finnish nominal declension classes:[15]

Class	Nom.sg	Gen.sg	Part.sg	Part.pl	Iness.pl	
4	*lasi*	*lasin*	*lasia*	*laseja*	*laseissa*	'glass'
9	*nalle*	*nallen*	*nallea*	*nalleja*	*nalleissa*	'teddy'
8	*ovi*	*oven*	*ovea*	*ovia*	*ovissa*	'door'
32	*kuusi*	*kuusen*	*kuusta*	*kuusia*	*kuusissa*	'six'
10	*kuusi*	*kuuden*	*kuutta*	*kuusia*	*kuusissa*	'spruce'

Shaded cells uniquely identify class assignment, while plain wordforms do not.

    d. Predictive wordforms identify correct class assignment

Stimulus: *tuohta* birchbark.part.sg:

kuusta : *tuohta* :: *kuusi* : *TUOHI*
Correct assignment to class 32

    e. Neutralized wordform: underdetermines correct class assignment

Stimulus: *nuken* puppet.gen.sg

nallen : *nuken* :: *nalle* : *NUKKE*
oven : *nuken* :: *ovi* : *NUKKI*
Correct class assignment to 9 indeterminate with respect to 9 v. 8

[14] We follow the basic representations and lines of argumentation in Paunonen 1976, Thymé 1993, and Thymé, Ackerman, and Elman 1994.

[15] 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.

Stimulus: *nukkeja* puppet.part.pl

n*alleja* : *nukkeja* :: *nalle* : *NUKKE*
laseja : *nukkeja* :: *lasi* : *NUKKI*
Correct class assignment to 9 indeterminate with respect to 9 v. 4

Stimuli: *nuken* puppet.gen.sg & *nukkeja* puppet.part.pl

Conjunction of wordforms is predictive – correct class assignment to 9

19. Strategy for pattern assignment
Paradigmatic patterns: The wordform in a specific cell or wordforms in several separate cells (i.e., patterns of cells) are diagnostic of declension class membership.[16]
20. This is (essentially) Stump and Finkel's notion of *dynamic* principal parts, contrasting with *static* and *adaptive* analyses.
    a. In fact, there are many equally good alternative sets of principal parts for Finnish, and many more solutions that are almost as good.
    b. We speculate that this is a common feature of complex morphological systems (cf. *resilience* in biological systems)
    c. A consequence: the *paradigm cell filling problem* is related to, but not quite the same as, the *class assignment problem*
21. General hypothesis of (sub)paradigm organization
Identifiable patterns of relatedness between wordforms in paradigms facilitate paradigm cell filling. Related wordforms are partitioned into (sub)paradigms with their own small systems of relatedness among forms.
What Finnish (sub)paradigms share are recurring formal elements , e.g., *lasi* occurs in nom. sg., gen. sg. & part. sg. while *lase* occurs in part. pl. and iness. pl.[17]
22. Paradigm Cell Filling problem: General formulation
    a. Given a lexeme L associated with a set of morphosyntactic properties (=morphosyntactic or grammatical word) and expressed by a surface wordform (exponent), 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?
    b. So, 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, i.e., given a novel inflected word form, what are all the other wordforms in its inflectional (and derivational) families?

[16] There is an alternative strategy for pattern assignment that will be largely ignored here. This is referred to as *Token analogy* in Thymé et. al. On this strategy class membership is deduced on the basis of phonological similarity of a new form to a previous form, but since phonologically similar wordforms may belong to different classes, such similarity may lead to erroneous classification of the novel word. See Thymé 1993 for Finnish, Pertsova 2004 for Russian, Albright and Hays 2002 on phonological neighborhoods.

[17] 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.

23. The importance of surface words
    Q: What forms the basis for prediction?
    A: Surface words and patterns of relatedness among surface words[18]

24. *UG Hypothesis of wordhood*[19]
    The natural domain of generalization for UG in morphology is the surface word as the exponent of the *grammatical/morphosyntactic word* expressed either synthetically or periphrastically (see Tundra Nenets dual local case relations), i.e. surface wordforms as the occupants of cells in paradigms.[20]

25. *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., "recurrent partials" such segments, tones, etc.
    For any given language: What are the (sub)patterns of (inter)predictability and what are the elements relevant to (inter)predictability?

26. Present task: (partial) Tundra Nenets (Samoyed) nominal inflection
    Given any Tundra Nenets inflected noun wordform, what are the remaining 209 forms of this lexeme for the morphosynactic feature property combinations CASE { nom, acc, gen, dat, loc, abl, pro}, NUMBER {singular, dual, plural}, POSSESSOR {3 pers. x 3 num.}? (7 x 3) + (7 x 3 x 3 x 3) = 210

a)	Stimulus:	Target	vs.	b)	Stimulus	Target
	nganuqmana	*ngano*			*wíngoqmana*	*wíh*
	boat.PL.PROS	boat.SG.NOM			tundra.PL.PROS	tundra.SG.NOM

IV. Patterns of predictability

Identifying patterns of (inter)predictability for a subset of Tundra Nenets nominal declensions within and across (sub)paradigms

27. Patternment within and across stem types – Absolute declension (=Non-possessive)
Lexical categories are divisible into the following gross stem type classification (ignoring the relevance of syllabicity, see Salminen (1997, 1998) for careful of exposition of Types and see VI below for use of these classes):

---

[18] 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).

[19] The operative notion of UG here is that characterized in Ackerman and Webelhuth 1998 as "grammatical archetypes". These are recurring constructs of languages that admit of grammatical generalizations and is orthogonal to claims about their innate versus emergent status.

[20] 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 e.g., 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 is non-nasalizing/devoicing glottal stop[21] (= q)

Type 1: polysyllabic vowel stem: *ngano* 'boat'

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

Type 2i: nasalizing glottal stem: *wíh/wíng* 'tundra'

	Singular	Plural	Dual
Nominative	*wíh*	*wíq*	*wíngh*
Accusative	*wímh*	*wíngo*	*wíngh*
Genitive	*wí^{0}h*	*wíngoq*	*wíngh*
Dative-Directional	*wíndh*	*wíngq*	*wíngh nyah*
Locative-Instrumental	*wíngana*	*wíngaqna*	*wíngh nyana*
Ablative	*wíngad*	*wíngat*	*wíngh nyad*
Prolative	*wímna*	*wíngoqmana*	*wíngh nyamna*

21 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 table, all of the statistical calculations are based on Salminen's phonological transcriptions of words.

Type 2ii: nasalizing glottal stem: *myaq/myad* 'hut'

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

28. A note on allomorphy: the occurrence of a specific allomorph, e.g., *wíngana*, where *-gana* (part of the family with *-xana*, *-kana*) leads to the inference that this word belongs to the class of stem final nasalizing glottals.
29. Basic observation about Tundra Nenets:
Nominal paradigms for all stem classes are partitioned into subparadigms each of which is defined by the presence of a characteristic and recurring stem, e.g., *ngano*, *nganu*, *nganoxo*
30. Hypothesis about the organization of Tundra Nenets paradigms
Subparadigms are domains of interpredictability among wordforms, rather than of derivability from a privileged base.[22]

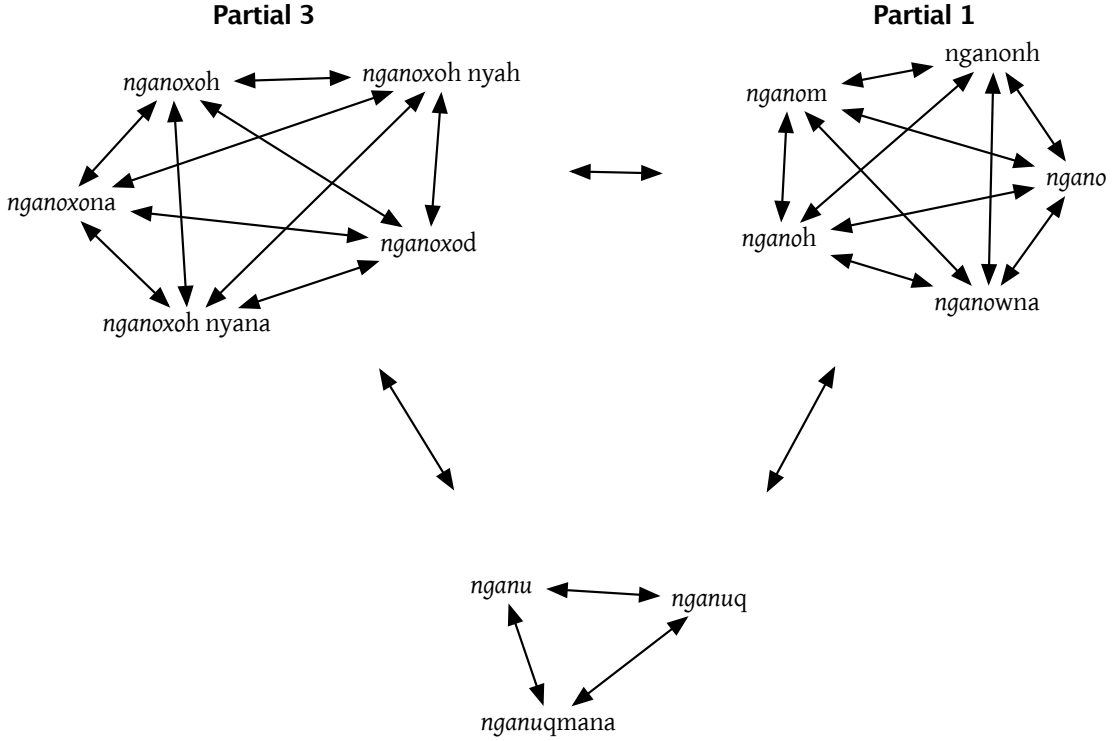
V. Competing hypotheses

Recurrent parts versus derivational bases

31. Bochner (1993): no form serves as privileged base
"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)
    a. Alliances of wordforms share formal properties, but the elements in such alliances need not be thought of as bearing derivational relations to one another, let alone to a single isolable base form.
    b. Symmetric relations among members of (sub)paradigms

[22] 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.)

c. The absence of privileged single base does not entail that there cannot be multiple subparadigms in which a particular recurring form (a *partial)* serves a pivotal role.



d. Each form in the subparadigm provides information about other forms in the subparadigm, i.e., members of subparadigms share partials.
"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)..." (Anderson 1985:53)

32. Albright (2002): a single privileged form serves as base
"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."[23] (Albright 2002:118)

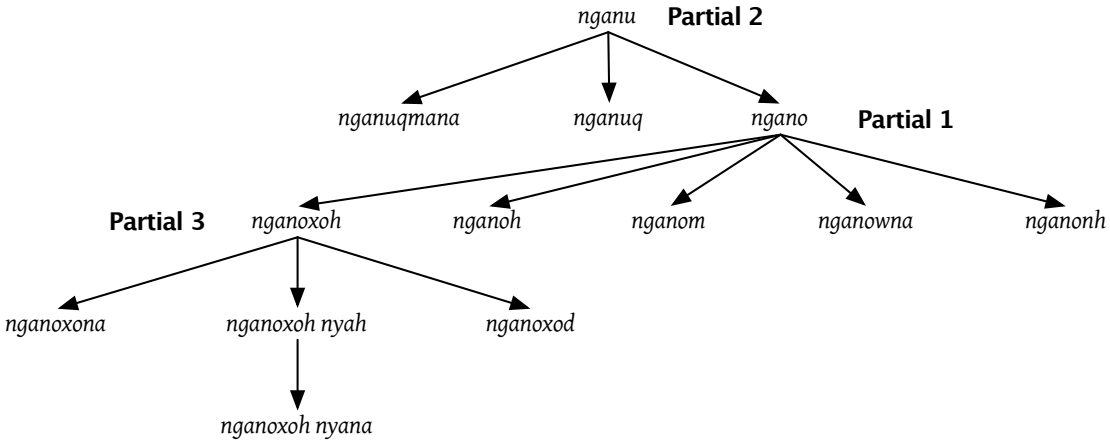
a. There is a single most informative surface form within any paradigm from which all other forms are derivable.

b. This single base form is the same for all classes. If the part. sg. , is identified as the base for declension class 1, then the part sg. form must likewise serve as the base for all other declension classes.

c. Asymmetric derivational relation between base and derived forms

[23] 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.)."

d. The postulation of a single surface base does 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)



e. Each local base, however, is similarly asymmetric in the relations between wordforms, namely, each subparadigm will contain a base from which the rest of the forms in it are derived.

f. Base gives information about derivate, but derivate does not give information about base.

33. The Goldilocks gambit: Looking for what's just right
If there's symmetry, fine, and if there is asymmetry, that's fine too.

34. Working hypothesis: Tundra Nenets is symmetrical
Tundra Nenets nominal paradigms are organized around several subparadigms, but the domains in which these bases play a role are domains of interpredictability of forms, rather than of derivability from a privileged base.

35. The basic picture:

a. (Sub)paradigms are systems of related wordforms organized around recurring partials.

b. Two of the three primary partials happen to be independent wordforms in Nenets, but they needn't be.

c. 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).

36. How can we compare these proposals?
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, i.e., test the Single Surface Base Hypothesis.[24]

[24] 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.

VI. Relative predictiveness of nom. sg. v. acc. pl.

A pilot study

37. Research question
    Between the nom. sg. and the acc. pl., which, if either, is more useful for predicting the other?

38. Data
    Corpus of 4,334 nominals extracted from Salminen's superb dictionary corpus of 16,403 entries based on Tereshchenko 1965 and specifying meaning, frequency, as well as the stem class assignment of entries

39. Consider the following pairs of nom.sg. and acc, pl. forms of related lexemes[25]

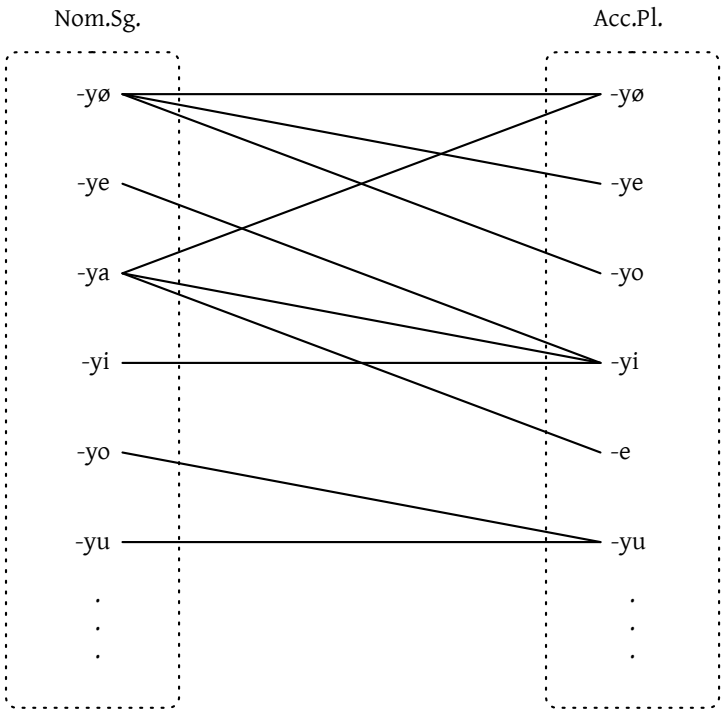
Nom sg.	Acc. pl.	Gloss
ngøno	*ngønu*	'boat'
lyabtu	*lyabtu*	'harnessed deer'
ngum	*nguwo*	'grass'
xa	*xawo*	'ear'
nyum	*nyubye*	'name'
yí	*yíbye*	'wit'
myir	*myirye*	'ware'
wíh	*wíngo*	'tundra'
weh	*weno*	'dog'
nguda	*ngudyi*	'hand'
xoba	*xob°*	'fur'
saw°nye	*saw°nyi*	'magpie'
tyírtya	*tyírtya*	'bird'

    a. Appearance of indeterminacy or uncertainty in both directions.
    b. From the perspective of a single base hypothesis the question is whether one is derivable from the other (or whether some 3rd form can serve as a base for both[26]),
    c. Given the maximal difference between these two forms we simply consider directionality of derivation between these two forms, ignoring other related wordforms.

40. Representative mappings between Nom. sg. and Acc. pl.

[25] These wordforms are taken from Salminen 1997 and, consequently, presented in their original transcription where a superscripted ° designates a schwa and ∅ 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.

[26] A comprehensive examination of Albright's informativeness measure would test the entire Tundra Nenets paradigm in order to identify the most informative form. It would also explore whether deviations of predictability from a single base are sensitive to phonological cues that we many not have not coded for.



41. The measure of predictability: *Entropy*
    a. Quantify "prediction" as a change in uncertainty, or 'information entropy' (Shannon 1948).
    b. Intuitively, the more choices there are, and the more evenly distributed their probabilities, the greater the uncertainty or surprise (on average) that a particular choice will be made, and the greater the entropy.
    c. Conversely, random variables with only a few possible outcomes, or with one or two highly probable outcomes and lots of rare exceptions, have a low entropy.
    d. Given a random variable X (which can take on one of a set of discrete choices) and their probabilities $P(X)$, we can calculate the entropy $H(X)$ of the variable.
42. Measuring entropy among Tundra Nenets nominal types
    a. 31 different types of nom. sg. nouns. Some, like the class of words ending in the reduced vowel *-ø*, are quite common, while others are quite rare.
    b. Overall, the entropy of this distribution H(nom.sg.) is 3.28 bits, i.e., it would take a minimum of 3.28 bits per word (on average) to list the nom.sg. types in a computerized dictionary.
    c. Similarly, 35 different types of accusative plurals, and the entropy H(acc.pl.) is 3.36 bits.
43. Calculating predictability: the size of the surprise
    a. Having quantified the degree of uncertainty in the choice of nom. sg. and acc.pl. types individually, we can now calculate "predictability".

b. Measure predictability using *conditional entropy* $H(Y|X)$, i.e., the uncertainty in the value of Y given that we already know the value of X.

c. 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.

44. Nom. sg. → Acc. pl (= H(acc.pl.|nom.sg.) vs. Acc. pl. → Nom. sg. (= H(nom.sg.|acc.pl.)

a. Consider H(acc.pl.|nom.sg.), i.e., the uncertainty in the acc. pl. given the nom. sg. There are a total of 52 nom.sg./acc.pl. combinations. In some cases, knowing the nom. sg. of a word uniquely identifies its acc. pl., e.g. the nom. sg. ending in *–ye* is always *-yi* in the acc. pl. With this type of nom.sg. word, there is no uncertainty in the acc.pl. and the conditional entropy H(acc.pl.|*-ye*)=0. 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 *-ø, -yi, -yø,* or *-e*.

b. 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 known to be needed to encode the acc.pl. Going in the other direction, from acc. pl. to nom. sg., the conditional entropy H(nom.sg.|acc.pl.)=0.51. In other words, the acc. pl. "predicts" all but 0.51 of the 3.28 bits in the nom. sg.

45. Interpretation of results

a. Acc.pl. is slightly more helpful for predicting the nom.sg. than vice versa, i.e., the conditional entropy is closer to 0 in the former than the latter.

b. But, the real conclusion is that neither is especially useful for predicting the other, since there's still plenty of surprise in the outcome.

46. Type versus token frequencies

a. The above calculations are based on *type* frequencies: each distinct noun lexeme is counted once.

b. We can also calculate conditional entropy on the basis of *token* frequencies: all occurrences of nouns in the corpus (a given type can be represented by several tokens)

c. Using token counts from Salminen's corpus, we get:

H(nom.sg.)	=	3.81 bits
H(acc.pl.)	=	4.02 bits
H(acc.pl.|nom.sg.)	=	0.91 bits
H(nom.sg.|acc.pl.)	=	0.70 bits

d. As before: neither form predicts the other very well.

47. Relatively low likelihood of encountering the acc. pl.
Frequency distributions of absolute declension for all case and number encodings of the 12,152 noun tokens in Salminen's sample sentence corpus :[27]

27 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

a. Nom. sg. represents 33.8% of the tokens, acc. pl. represents 2.7%.
b. If the system were organized around the need to encounter the acc. pl., paradigm cell filling would be hampered, given the low probability of encountering this form.
c. The situation is even worse for Partial 3, based on the direct case dual forms (0.1% of the tokens).
d. In fact, this is likely true for all individual wordforms (except the nom. sg. and the gen.sg.)

48. There is a much higher likelihood of encountering the partials associated with acc. pl. if we posit subparadigms.
Frequency distributions for all absolute and possessive forms:

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

a. Cues for wordshapes can be gotten even for lower frequency forms if we assume alliances of related forms within subparadigms, i.e., any form within a subparadigm predicts the others and may provide clues for forms in other subparadigms.
b. Note that even if a derivability relation had been identifiable, this would not have accounted for the evident subpatterns of shared forms as described above, rendering such subpatterns epiphenomenal, rather than central to organization. (we have no explanation for why the alliances consist of the forms and feature sets they do.)

49. Summary of results
a. Both type and token calculations suggest that, for the comparison of nom. sg. and acc. pl., neither reliably serves as the single base from which the other is predicted.
b. These equivocal results with respect to directionality of prediction, contrast with the overwhelming likelihood of encountering nom. sg. versus acc. pl. on the basis of frequency distributions.

c. 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's important that the aggregate frequency of partials be high enough to be useful.

VII. Provisional conclusions and ramifications

50. Bochner's symmetrical pattern sets and Albright's asymmetric local bases are both used to model paradigm structure
    a. The two models make very different predictions when considered in the light of the paradigm cell filling problem.
    b. In Albright's model, derived forms should be predictable from bases, but there is no reason to expect bases to be predictable from derived forms or derived forms to be predictable from each other.
    c. Bochner's model, on the other hand, allows for potentially complex interrelations between forms in the same paradigm or subparadigm.
    d. (Sub)paradigms are organized in terms of patterns of whole word relatedness with members of (sub)paradigms exhibiting interpredictability: this facilitates solving the paradigm cell filling problem, i.e., to reliably predict an inflected form of a word given any other inflected form in languages like Tundra Nenets.
    e. Note that we've focused here on the role analogy plays in organizing wordforms in synchronic paradigms, and not to the role that analogy plays in driving changes in paradigms.
51. Ramifications: a developmental issue concerning learnability
    a. How do children go about identifying the relevant dimensions of morphosyntactic properties and how do they isolate the appropriate patterns of surface exponence?
    b. The whole word hypothesis is consistent with a learning strategy that begins with what children are exposed to concretely and then discovers/develops relatedness schemata of increasing abstractness which license inferences about novel wordforms (Tomasello 2003; Gentner and Namy 2004; Pinker 1984; MacWhinney 1978, among others.)
    c. So, 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.
52. Future prospects
    a. Learning as dimensionality reduction
    b. Map alliances among forms by extending information theoretic measures to beyond nom.sg./acc.pl. to complete Tundra Nenets paradigms and to paradigms in other morphologically complex languages (Chaha (Semitic), Moro (Kordofanian))
    c. Explore influences on learnability via connectionist modeling of 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 20th century.* University of Chicago Press.

Anderson, S.R. 1992. *A-morphous morphology.* Cambridge University Press.

Aronoff, M. 1994. *Morphology by itself.* MIT Press.

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,

Cech, P. 1995/1996. Inflection/derivation in Sepečides-Romani. *Acta Lingustica Hungarica* 43.1–2:67–92

Clahsen, H. 1999. Lexical entries and rules of language: A multidisciplinary study of German inflection. *Behavioral and Brian 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.

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.

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*]. 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.

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.

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.

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.

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*.

Tereshchenko, N. M. 1965/2003 *Nenetsko-ruskii slovar'* [*Nenets-Russian dictionary*]. St. Petersburg.

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

Appendix: Possessive declension (partial paradigm)

		1st sing	2nd sing	3rd sing
SING	N	*nganow*	*nganor*	*nganoda*
	A	*nganow*	*nganomd*	*nganomda*
	G	*nganonyi*	*nganond*	*nganonda*
	D	*nganoxonyi*	*nganoxond*	*nganoxonda*
DUAL	N	*nganoxoyunyi*	*nganoxoyud*	*nganoxoyuda*
	A	*nganoxoyunyi*	*nganoxoyud*	*nganoxoyuda*
	G	*nganoxoyun(yi)*	*nganoxoyut*	*nganoxoyuta*
	D	*nganoxoqn(yi) nyah*	*nganoxoyut nyah*	*nganoxoyuta nyah*
PLUR	N	*nganunyi*	*nganud*	*nganuda*
	A	*nganunyi*	*nganud*	*nganuda*
	G	*nganuqn(yi)*	*nganut*	*nganuta*
	D	*nganoxoqn(yi)*	*nganoxot*	*nganoxota*
		1st du	2nd du	3rd du
SING	N	*nganomyih*	*nganoryih*	*nganodyih*
	A	*nganomyih*	*nganomdyih*	*nganomdyih*
	G	*nganonyih*	*nganondyih*	*nganondyih*
	D	*nganoxonyih*	*nganoxondyih*	*nganoxondyih*
DUAL	N	*nganoxoyunyih*	*nganoxoyudyih*	*nganoxoyudyih*
	A	*nganoxoyunyih*	*nganoxoyudyih*	*nganoxoyudyih*
	G	*nganoxoyunyih*	*nganoxoyutyih*	*nganoxoyutyih*
	D	*nganoxoyunyih nyah*	*nganoxoyutyih nyah*	*nganoxoyutyih nyah*
PLUR	N	*nganunyih*	*nganudyih*	*nganudyih*
	A	*nganunyih*	*nganudyih*	*nganudyih*
	G	*nganuqnyih*	*nganutyih*	*nganutyih*
	D	*nganoxoqnyih*	*nganoxotyih*	*nganoxotyih*
		1st pl	2nd pl	3rd pl
SING	N	*nganowaq*	*nganoraq*	*nganodoh*
	A	*nganowaq*	*nganomdaq*	*nganomdoh*
	G	*nganonaq*	*nganondaq*	*nganondoh*
	D	*nganoxonaq*	*nganoxo ndaq*	*nganoxondoh*

DUAL	N	*nganoxoyunaq*	*nganoxoyudaq*	*nganoxoyudoh*
	A	*nganoxoyunaq*	*nganoxoyudaq*	*nganoxoyudoh*
	G	*nganoxoyunaq*	*nganoxoyutaq*	*nganoxoyutoh*
	D	*nganoxoyunaq nyah*	*nganoxoyutaq nyah*	*nganoxoyutoh nyah*
PLUR	N	*nganunaq*	*nganudaq*	*nganudoh*
	A	*nganunaq*	*nganudaq*	*nganudoh*
	G	*nganuqnaq*	*nganutaq*	*nganutoh*
	D	*nganoxoqnaq*	*nganoxotaq*	*nganoxotoh*