



Deseret Language and Linguistic Society Symposium

Volume 7 | Issue 1

Article 9

3-27-1981

Probabilistic Descriptions of Finnish Morphology

Royal Skousen

Follow this and additional works at: <https://scholarsarchive.byu.edu/dlls>

BYU ScholarsArchive Citation

Skousen, Royal (1981) "Probabilistic Descriptions of Finnish Morphology," *Deseret Language and Linguistic Society Symposium*: Vol. 7 : Iss. 1 , Article 9.

Available at: <https://scholarsarchive.byu.edu/dlls/vol7/iss1/9>

This Article is brought to you for free and open access by the Journals at BYU ScholarsArchive. It has been accepted for inclusion in Deseret Language and Linguistic Society Symposium by an authorized editor of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu, ellen_amatangelo@byu.edu.

PROBABILISTIC DESCRIPTIONS OF FINNISH MORPHOLOGY

Royal Skousen
Brigham Young University

In this paper I use a general probabilistic method for describing linguistic behavior. In this approach I first define the CERTAINTY of a rule as the probability that two randomly selected occurrences of the rule have the same outcome. The certainty of a system of rules is then defined as the weighted sum of the certainties of each rule in the system. Given these definitions, it can be shown that a CORRECT description is a system of rules for which certainty is maximized. Each rule in a correct description is HOMOGENEOUS in behavior (that is, all the subrules of a homogeneous rule behave no differently from the rule). Finally, I define an OPTIMAL description as a structured system of rules that uses the least number of rules to correctly describe behavior. Optimal descriptions can be used (1) to classify regularities according to their-degree of similarity, (2) to distinguish exceptional behavior from regular behavior, and (3) to predict the order in which rules should be learned so as to minimize uncertainty as quickly as possible.

The main purpose of this paper is to apply this method to several different types of morphological alternation in the verbal system of Finnish. In particular, I will deal with two questions: (1) What is the optimal presentation of Finnish morphological relations so that a person learning the language can minimize morphological errors as quickly as possible? (2) Which inflectional forms are the most helpful in predicting other inflectional forms?

In order to answer these questions from a probabilistic point of view, we must consider the frequency of Finnish words. In this paper estimates of word frequency are based on Pauli Saukkonen's Suomen kielen taajuus sanasto. As a consequence, infrequently occurring verbs play only a minor role in the analysis. In addition, it should be noted that Saukkonen gives only the frequencies for separate lexical entries, so that the frequencies of individual forms are not directly determinable. Saukkonen's statistics can be used to determine the frequencies of the verb paradigms in Finnish, but these statistics may not always reflect the relative frequencies of certain inflected forms -- especially when an inflected form is not fully productive. For this reason, I will deal in this paper with only fully productive inflectional forms.

To begin with, let us consider how we might optimally predict the imperfect stem from the present stem. There are nine different alternations between the present stem and the imperfect. Given these possible outcomes, we split up the contextual space of the present stem into as few contexts as possible, simultaneously requiring that each context be homogeneous in behavior. In other words, our description must be correct. For each context we specify the alternation that occurs most

frequently in that context -- that is, we do not attempt to predict the less frequent alternations that a word might take. For instance, although lähteä has two possible imperfect forms (lähti and läksi), the rules in this paper will account for only the most frequent form, lähti.

The optimal contextual space can be best represented in terms of a Venn diagram:

PRESENT STEM → IMPERFECT

<u>context</u>	<u>alternation</u>	<u>freq.</u>	<u>rank</u>	<u>examples</u>
-C[unround V] 	V-i	57951	1	muistaa, lukea, toimia
	a-oi	3744	4	kasvaa, antaa
	tV-si	66	14	taitaa
	tV-si	1890	7	myöntää, ymmärtää, sisältää
	V-i	11	16	yltää
	V-i	8	17	kyntää
	V-i	3	18	sortaa
	tV-si	411	11	tuntee
	tV-si	419	10	tietää
	tV-si	177	12	löytää
tV-si	102	13	pyytää	
tV-si	60	15	huutaa	
-C[round V]	∅-i	14353	2	sanoa, kuulua
-V ^ä	A-si	5832	3	huomata, pudota, poiketa
-V ^í :	V:-Vi	2760	5	saada, jäädä
-Vi	=	2751	6	uida, voida
-[high V][mid V]	V _x V _y -V _y i	800	8	tuoda, viedä
käy	äy-ävi	467	9	käydä

(cf. appendix at end of paper for explanation of symbols)

This diagram tells us several important facts about the forms of the Finnish imperfect. First, the general contextual space is split into seven main rules, six of which are general rules and one is based on a single lexical form, käydä. Main rules are represented by rectangles, exceptions by circles. Only the first of these main rules is non-homogeneous -- that is, it is the only one that contains subrules that behave differently from the rule itself. Of the seven exceptional rules to this first rule, five are lexical and two are general. And both these general exceptions also contain lexical exceptions. In all, there are 18 rules in the optimal description of the imperfect. We refer to the total number of rules in a description as its complexity.

Comments: (1) The context $-\hat{a}(V)C_1a$ occurs instead of the more general context $-[unround V](V)C_1a$ since no examples like virkkaa (virkka-virkkoi) are found in Saukkonen. (2) The general exception $-\hat{a}(V)C_1a$ takes precedence over the $-S_d tA$ exception (e.g. antaa (anta-antoi)¹ rather than anta-**ansi*). This preference is represented in the Venn diagram by letting the preferred context overlap the other.

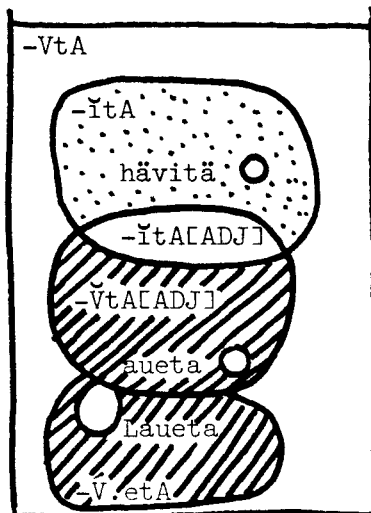
Finally, we may rank the rules of the description in terms of frequency. This ranking defines the order in which these rules should be learned so as to minimize error as quickly as possible.

One important problem in any morphological analysis is to determine which inflectional forms are the most helpful in predicting other inflectional forms. For example, consider the question of whether it is easier to predict the present stem from the infinitive form (the dictionary form) -- or, vice versa, to predict the infinitive form from the present stem. Let us first determine the complexity of predicting the present stem from the infinitive. If we ignore for the moment the problem of gradation, we see that there are nine possible alternations. The optimal description for predicting the present stem from the infinitive contains 14 rules:

INFINITIVE → PRESENT STEM

<u>context</u>	<u>alternation</u>	<u>freq.</u>	<u>examples</u>
-V \check{A}	A- \emptyset	43877	sanoa, lukea, toimia
$-S_d S_d A$	$S_d S_d A - S_d e$	29830	mennä, purra, vaihdella
-V dA	$dA - \emptyset$	6778	uida, saada, arvioida
CC[front \acute{V}]hdä	hdä-ke	1563	tehdä, nähdä
-stA	stA-se	1279	päästä, nousta
⊗ CC[high V][mid V]stA	stA-kse	64	juosta, syöstä

(diagram continues on next page)



tA-A	5611	haluta, pudota
tA-tse	2122	mainita, ansaita
tA-A	68	hävitä
tA-A	121	selvitä, lämmitä
tA-ne	374	väheta, edeta, rohjeta
tA-A	22	aueta
tA-ne	7	laueta, raueta
tA-ne	100	kyeta, paeta, vaeta

Comment: The symbol [ADJ] means that the verb is morphologically derived from an adjective (e.g. pienetä from pieni and selvitä from selvä). However, for some verbs this morphological relationship may be obscure, especially when the corresponding adjective is infrequent or the phonetic relationship is unclear: e.g. loitota (from loitto), huveta (from hupa), or edeta (from the stem ete- [cf. eteen, edellä]).

But the main source of difficulty in predicting the present stem from the infinitive involves the question of gradation. We are required in certain contexts to predict the strong grade from the weak grade -- that is, we need to use what might be called (from a historical point of view) REVERSE GRADATION.

REVERSE GRADATION.

[voiceless stop] - [geminate voiceless stop]	mitata-mittaa
[geminate sonorant] - [sonorant][homorganic stop]	verrata-vertaa
d - t	todeta-totea
Ŵ.V - Ŵ.kV	huoata-huokaa
Ŵ: - ŴkV	koota-kokoa
Ŵhje - Ŵhke	puhjeta-puhkea
ŴLje - ŴLke	herjetä-herkeä
ŴhV - ŴhkV	uhata-uhkaa
ŴLV - ŴLkV	pelätä-pelkää

Amazingly, we can predict the gradation very well. Only two places cause a lot of difficulty:

<u>context</u>	<u>gradation</u>	<u>freq.</u>	<u>examples</u>
U	no	80608	joutua, kokeilla, pysäköidä
-VtA	yes	3549	upota, langeta, koota
-V̄tA[ADJ]	yes*	517	hiljetä, rohjeta, levitä
-itA	no	2122	hallita, kyyditä, harkita
-vV̄tA	no	497	korvata, siivota, avata
-vetA	yes	91	kiivetä, ruveta, hävetä
○ kivetä ● 8 exceptions	no	2	kivetä
-V̄LVtA ● 5 exceptions	yes	288	tavata, kaivata, luvata, levätä, kelvata, evätä, turvota, kivuta
-V̄hVtA	no	995	haluta, seurata, palata
○ uhata	yes	161	pelätä, karata, hylätä, keretä, perata
○ tallata, narrata	no	154	kohota, vihata, tuhota
○ niitata	no	38	uhata
-C V̄llA ○	yes	9	tallata, narrata (cf. verrata, vallata)
-vëllA	no	2	niiata (cf. huoata, seota)
-V̄LjellA	yes	2572	kuunnella, jaella, puhalla, kierrellä
○	no	163	arvella, kävellä, palvella
○	no	32	varjella, viljellä
● rangaista	yes	16	rangaista (cf. aukaista, natista)

complexity: 30

* In those cases where we depend upon a morphologically related adjective to predict the form of the present stem, we can also use the adjective form to predict the gradation: e.g. rohjeta-rohkene (cf. rohkea); levitä-leviä (cf. leveä).

Thus the total complexity in predicting the present stem from the

infinitive is $14 + 30 = 44$.

On the other hand, predicting the infinitive from the present stem is really quite simple:

PRESENT STEM → INFINITIVE

<u>context</u>	<u>alternation</u>	<u>freq.</u>	<u>examples</u>
-C \check{V}	∅-A	43877	sanoa, kuulua
-VS _d e	S _d e-S _d S _d A	29830	mennä, kuunnella, purra, tulla
-Vne	ne-tA	482	parata, pienetä
-itse	tse-tA	2122	mainita, ansaita
C[front V]ke	ke-hdä	1563	nähdä, tehdä (cf. hakea, kokea, lukea, tukea, pukea)
-Vse	se-stA	1279	nousta, aukaista
C[high V][mid V]kse	kse-stA	64	juosta, syöstä
-VV	∅-dA	6778	saada, tuoda, uida
-V \check{A}	A-tA	5821	huomata, pudota

complexity: 9

(Note: exceptions to the first main rule all end in e)

GRADATION

<u>context</u>	<u>gradation</u>	<u>freq.</u>	<u>examples</u>
U	no	82925	merkitse-merkitä, aukaise-aukaista, epäile-epäillä
-C ₀ VS _d e	yes	3054	kuuntele-kuunnella, kykene-kyetä
-V \check{A}	yes	5821	putoa-pudota, tapaa-tavata
rangaista	yes	16	rankaise-rangaista

complexity: 4

Thus the total complexity in predicting the infinitive from the present stem is $9 + 4 = 13$. We therefore conclude that the present stem of the

verb is much more helpful in predicting the inflectional verb forms of Finnish than the infinitive is.

Sometimes it is advantageous to consider more than one base in predicting another form. Consider, for instance, the conditional stem. If we are restricted to using only one form to predict the conditional, the best form would be the present stem. The present stem requires seven rules to describe the formation of the conditional stem:

PRESENT STEM → CONDITIONAL

<u>context</u>	<u>alternation</u>	<u>freq.</u>	<u>examples</u>
$-.C_0\check{A}$	\emptyset -isi	26210	muista-muistaisi, selviä-selviäisi
-i	\emptyset -si	5646	ui-uisi, toimi-toimisi
- \check{e}	e-isi	37662	luke-lukisi, tunte-tuntisi
-[round \check{V}]	\emptyset -isi	14368	sano-sanoisi, kuulu-kuuluisi
käy	äy-ävisi	467	käy-kävisi
-V:	V:-Visi	6663	pelkää-pelkäisi, myy-myisi, saa-saisi
-[high V][mid V]	$V_x V_y -V_y$ isi	800	tuo-toisi, vie-veisi


But if the present stem does not end in a stressless low vowel, then the conditional stem can be simply derived from the imperfect stem by adding si. If we take advantage of this close relationship between the imperfect and the conditional stem, we get a description of the conditional which involves only four rules:

PRESENT STEM/IMPERFECT → CONDITIONAL

- (1) If the present stem ends in a stressless low vowel, use the present stem as the base form:

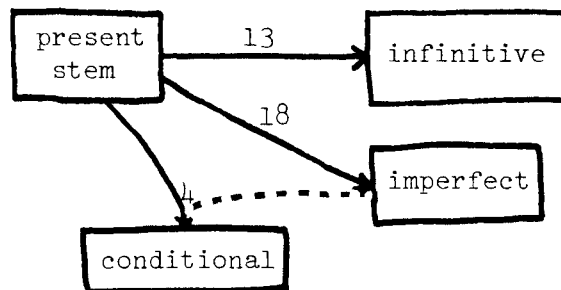
$-.C_0\check{A}$	\emptyset -isi	26210	muista-muistaisi, halua-haluaisi
- \check{A} :	A:-Aisi	3903	huomaa-huomaisi, avaa-avaisi

- (2) If the present stem does not end in a stressless low vowel, use the imperfect stem as the base form:

U	∅-si	61292	koske-koski-koskisi, saa-sai-saisi, vie-vei-veisi, käy-kävi-kävisi, sano-sanoi-sanoisi, toimi-toimi- toimisi
tunsi 	si-tisi	411	tunte-tunsi-tuntisi

There is some evidence for this relationship between the conditional and the imperfect. Consider, for instance, the occurrence of the analogical form läksisi in dialects where the historical läksi has been retained. Similarly, some speakers of Finnish accept tunsi as a possible conditional form for tuntea. The form tuntisi is exception only when derived from the imperfect.

By combining these results, we derive a system of optimal relationships for the Finnish verb. We observe the primacy of the present stem in such a system.



APPENDIX: Explanation of Symbols

C	consonant	C_0	any number of consonants, including none
		C_1	at least one consonant
V	(short) vowel	V:	long vowel
		V_x, V_y	indexed vowels
A	(short) low vowel	(<u>a</u> or <u>ä</u>)	
S_d	dental sonorant	(<u>l</u> , <u>n</u> , or <u>r</u>)	
L	liquid	(<u>l</u> or <u>r</u>)	
ˈ	primary stress		
˘	unstressed		
.	syllable boundary	=	identity
()	optionality	∅	the null symbol