Probabilistic Descriptions of Finnish Morphology

Royal Skousen
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 kielten 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 läks), 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:

<table>
<thead>
<tr>
<th>context</th>
<th>alternation</th>
<th>freq.</th>
<th>rank</th>
<th>examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>-CUnround VJ</td>
<td>V-i</td>
<td>57951</td>
<td>1</td>
<td>muistaa, lukea, toimia</td>
</tr>
<tr>
<td>-a(V)C</td>
<td>a-oI</td>
<td>3744</td>
<td>4</td>
<td>kasvaa, antaa</td>
</tr>
<tr>
<td></td>
<td>tV-si</td>
<td>66</td>
<td>14</td>
<td>taitaa</td>
</tr>
<tr>
<td></td>
<td>tV-si</td>
<td>1890</td>
<td>7</td>
<td>myöntää, ymmartää, sisältää</td>
</tr>
<tr>
<td></td>
<td>V-i</td>
<td>11</td>
<td>16</td>
<td>yltää</td>
</tr>
<tr>
<td></td>
<td>V-i</td>
<td>8</td>
<td>17</td>
<td>kyntää</td>
</tr>
<tr>
<td></td>
<td>V-i</td>
<td>3</td>
<td>18</td>
<td>sortaa</td>
</tr>
<tr>
<td></td>
<td>tV-si</td>
<td>411</td>
<td>11</td>
<td>tuntea</td>
</tr>
<tr>
<td></td>
<td>tV-si</td>
<td>419</td>
<td>10</td>
<td>tietää</td>
</tr>
<tr>
<td></td>
<td>tV-si</td>
<td>177</td>
<td>12</td>
<td>löytää</td>
</tr>
<tr>
<td></td>
<td>tV-si</td>
<td>102</td>
<td>13</td>
<td>pyytää</td>
</tr>
<tr>
<td></td>
<td>tV-si</td>
<td>60</td>
<td>15</td>
<td>huutaa</td>
</tr>
<tr>
<td>-CRound VJ</td>
<td>Ø-i</td>
<td>14353</td>
<td>2</td>
<td>sanoa, kuulua</td>
</tr>
<tr>
<td>-VÄ</td>
<td>A-sI</td>
<td>5832</td>
<td>3</td>
<td>huomata, pudota, poiketa</td>
</tr>
<tr>
<td>-V:</td>
<td>V:-Vi</td>
<td>2760</td>
<td>5</td>
<td>saada, jäädä</td>
</tr>
<tr>
<td>-Vi</td>
<td>=</td>
<td>2751</td>
<td>6</td>
<td>uida, voida</td>
</tr>
<tr>
<td>-[high VJ][mid VJ]</td>
<td>x y V i</td>
<td>800</td>
<td>8</td>
<td>tuoda, viedä</td>
</tr>
<tr>
<td>käy</td>
<td>Æy-Ævi</td>
<td>467</td>
<td>9</td>
<td>käydä</td>
</tr>
</tbody>
</table>

(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 -ä(V)Cl.a occurs instead of the more general context -unround Vj(V)C.a since no examples like virkkaa (virkka-virkkoj) are found in Saukkonen. (2) The general exception -ä(V)C.a takes precedence over the -StA exception (e.g. antaa (anta-antoi) 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:

<table>
<thead>
<tr>
<th>context</th>
<th>alternation</th>
<th>freq.</th>
<th>examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>-Vää</td>
<td>A-∅</td>
<td>43877</td>
<td>sanoa, lukea, toimia</td>
</tr>
<tr>
<td>-SdA</td>
<td>SdA-∅</td>
<td>29830</td>
<td>mennä, purra, vaihdella</td>
</tr>
<tr>
<td>-VdA</td>
<td>dA-∅</td>
<td>6778</td>
<td>uida, saada, arvioida</td>
</tr>
<tr>
<td>CCfront VJhdä</td>
<td>hdä-ke</td>
<td>1563</td>
<td>tehdä, nähdä</td>
</tr>
<tr>
<td>-stA</td>
<td>stA-se</td>
<td>1279</td>
<td>päästä, nousta</td>
</tr>
<tr>
<td>STChigh VJmid VJstA</td>
<td>stA-kse</td>
<td>64</td>
<td>juosta, syöstä</td>
</tr>
</tbody>
</table>

(diagram continues on next page)
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 edetä (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 |
| ū.V - ū.kV | huoata-huokaa |
| ū: - ūkV | koota-kokoa |
| ūhje - ūhke | puhjeta-puhkea |
| ūLje - ūLke | herjetä-herkeä |
| ūhV - ūhkV | uhata-uhkaa |
| ūLV - ūLkV | pelätä-pelkää |
Amazingly, we can predict the gradation very well. Only two places cause a lot of difficulty:

<table>
<thead>
<tr>
<th>context</th>
<th>gradation</th>
<th>freq.</th>
<th>examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>no</td>
<td>80608</td>
<td>joutua, kokeilla, pysäköidä</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>3549</td>
<td>upota, langeta, koota</td>
</tr>
<tr>
<td></td>
<td>yes*</td>
<td>517</td>
<td>hiljetä, rohjeta, levitä</td>
</tr>
<tr>
<td></td>
<td>no</td>
<td>2122</td>
<td>hallita, kyyditä, harkita</td>
</tr>
<tr>
<td></td>
<td>no</td>
<td>497</td>
<td>korvata, siivota, avata</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>91</td>
<td>kiivetä, ruveta, hävetä</td>
</tr>
<tr>
<td></td>
<td>no</td>
<td>2</td>
<td>kivetä</td>
</tr>
<tr>
<td></td>
<td>yes*</td>
<td>288</td>
<td>tavata, kaivata, luvata, levätä, kelvata, evätä, turvota, kivuta</td>
</tr>
<tr>
<td></td>
<td>no</td>
<td>995</td>
<td>haluta, seurata, palata</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>161</td>
<td>pelätä, karata, hylätä, keretä, perata</td>
</tr>
<tr>
<td></td>
<td>no</td>
<td>154</td>
<td>kohota, vihata, tuhota</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>38</td>
<td>uhata</td>
</tr>
<tr>
<td></td>
<td>no</td>
<td>9</td>
<td>tallata, narrata (cf. verrata, vallata)</td>
</tr>
<tr>
<td></td>
<td>no</td>
<td>2</td>
<td>niïata (cf. huoata, seota)</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>2572</td>
<td>kuunnella, jaella, puhallella, kierrellä</td>
</tr>
<tr>
<td></td>
<td>no</td>
<td>163</td>
<td>arvella, kävellä, palvella</td>
</tr>
<tr>
<td></td>
<td>no</td>
<td>32</td>
<td>varjella, viljellä</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>16</td>
<td>rangaista (cf. aukaista, natista)</td>
</tr>
</tbody>
</table>

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 $\rightarrow$ INFINITIVE**

<table>
<thead>
<tr>
<th>context</th>
<th>alteration</th>
<th>freq.</th>
<th>examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>-CV</td>
<td>$\emptyset$-A</td>
<td>43877</td>
<td>sanoa, kuulua</td>
</tr>
<tr>
<td>-VS,e</td>
<td>S$_d$e-S$_d$A</td>
<td>29830</td>
<td>mennä, kuunnella, purra, tulla</td>
</tr>
<tr>
<td>ne-tA</td>
<td>482</td>
<td>parata, pienetä</td>
<td></td>
</tr>
<tr>
<td>tse-tA</td>
<td>2122</td>
<td>mainita, ansaita</td>
<td></td>
</tr>
<tr>
<td>ke-hdä</td>
<td>1563</td>
<td>nähdä, tehdä (cf. hakea, kokea, lukea, tukea, pukea)</td>
<td></td>
</tr>
<tr>
<td>se-stA</td>
<td>1279</td>
<td>nousta, aukaista</td>
<td></td>
</tr>
<tr>
<td>kse-stA</td>
<td>64</td>
<td>juosta, syöstä</td>
<td></td>
</tr>
<tr>
<td>-VV</td>
<td>$\emptyset$-dA</td>
<td>6778</td>
<td>saada, tuoda, uida</td>
</tr>
<tr>
<td>A-tA</td>
<td>5821</td>
<td>huomata, pudota</td>
<td></td>
</tr>
</tbody>
</table>

complexity: 9

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

**GRADATION**

<table>
<thead>
<tr>
<th>context</th>
<th>gradation</th>
<th>freq.</th>
<th>examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>no</td>
<td>82925</td>
<td>merkitse-merkitä, aukaise-aukaista, epällä-epällä</td>
</tr>
<tr>
<td>-CV,VS,e</td>
<td>yes</td>
<td>3054</td>
<td>kuuntele-kuunnella, kykene-kyetä</td>
</tr>
<tr>
<td>rangaista</td>
<td>yes</td>
<td>5821</td>
<td>putoa-pudota, tapaa-tavata</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>16</td>
<td>rankaise-rangaista</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>context</th>
<th>alternation</th>
<th>freq.</th>
<th>examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>-.C.ā</td>
<td>ő-isi</td>
<td>26210</td>
<td>muista-muistaisi, selviä-selviäisi</td>
</tr>
<tr>
<td>-i</td>
<td>ő-si</td>
<td>5646</td>
<td>ui-uiisi, toimi-toimisi</td>
</tr>
<tr>
<td>-ē</td>
<td>e-isi</td>
<td>37662</td>
<td>luke-lukisi, tunte-tuntisi</td>
</tr>
<tr>
<td>[-round ū]</td>
<td>ő-si</td>
<td>14368</td>
<td>sano-sanoisi, kuulu-kuuluisi</td>
</tr>
<tr>
<td>käy</td>
<td>äy-ävisi</td>
<td>467</td>
<td>käy-kävisi</td>
</tr>
<tr>
<td>-V:</td>
<td>V: -visi</td>
<td>6663</td>
<td>pelkää-pelkäisi, myy-myvisi, saa-saisi</td>
</tr>
<tr>
<td>[-[high V][mid V]]</td>
<td>V.V_V.visi</td>
<td>800</td>
<td>tuo-toisi, vie-veisi</td>
</tr>
</tbody>
</table>

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:

<table>
<thead>
<tr>
<th>context</th>
<th>alternation</th>
<th>freq.</th>
<th>examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>-.C.ā</td>
<td>ő-isi</td>
<td>26210</td>
<td>muista-muistaisi, halua-haluaisi</td>
</tr>
<tr>
<td>-ā:</td>
<td>ā: -āisi</td>
<td>3903</td>
<td>huomaa-huomaisi, avaa-avaisi</td>
</tr>
</tbody>
</table>

2. If the present stem does not end in a stressless low vowel, use the imperfect stem as the base form:
There is some evidence for this relationship between the conditional and the imperfect. Consider, for instance, the occurrence of the analogical form laksisi in dialects where the historical läksi has been retained. Similarly, some speakers of Finnish accept tunsisi 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  
  - $\hat{a}$ or $\overline{a}$

- **S_d** dental sonorant  
  - $\hat{1}$, $n$, or $r$

- **L** liquid  
  - $\hat{1}$ or $r$

- **.`** primary stress

- **'** unstressed

- **.** syllable boundary  
  - $=$ identity

- **( )** optionality

- **∅** the null symbol