Analogical Descriptions of Variation

Royal Skousen
The purpose of this paper is to compare two different approaches to describing linguistic variation.[1] The first of these approaches, commonly called structuralist, is the traditional method for describing behavior. Its methods are found in many diverse fields -- from biological taxonomy to literary criticism. A structuralist description can be broadly characterized as a classificatory system (normally defined as a system of rules). The fundamental question that a structuralist description attempts to answer is how a general contextual space should be split up. Structuralists have implicitly assumed that descriptions of behavior should not only be correct, but should also minimize the number of rules and permit only the simplest possible contextual specifications. It turns out that these intuitive notions can actually be derived from more fundamental statements about the uncertainty of rule systems.

Within a structuralist framework, we can identify three fundamental types of description according to how the contextual space is split up: (1) exceptional versus regular behavior; (2) categorical behavior; and (3) idiosyncratic behavior:

Problems arise when we try to use rules to predict behavior. A system of rules partitions the contextual space. This naturally implies a rule of usage which permits one and only one rule to apply to any given fully-specified context. When we use this rule of usage, the contextual partitioning forces an exact demarcation in predicted behavior as we move across the contextual space from one rule context to another. Consider the predicted behavior for our three fundamental types of description; in each case we get abrupt, distinct shifts in behavior:
Yet in all three of these cases, evidence from language behavior indicates that transitions across the contextual space are gradual and probabilistic:

As an example, consider the well-known results of perceptual tests between voiced and voiceless stops which show a gradual shifting towards the voiced stops as voicing onset time is increased:[2]

Or consider Labov's semantic experiment in which he found that as the relative width of a cup-like object is increased, the greater the chances speakers will identify the cup-like object as a bowl:[3]

Similarly, evidence from linguistic behavior shows that forms close to exceptional items may occasionally behave like those exceptions. Consider, for instance, repeated misspellings such as GREAD for grade,[4] which is apparently based on the exceptional spelling of the word great. A similar example is the spelling INCHOIR for inquire, obviously based on the uniquely-spelled word choir.[5] Johnson and Venezky have provided
examples of synthetic English words taking on exceptional pronunciations when these made-up words are close to certain exceptionally-spelled words. For example, PLOOD was frequently pronounced with the /ʌ/ vowel (as in the exceptionally spelled but very similar words blood and flood). Similarly, THEAT was frequently pronounced with the /ɛ/ vowel (as in the word threat, a very similar form).[6] The same effect has been observed in studies of word formation. For instance, I have recorded incorrect plural forms such as axen (for axes), which is obviously based on the exceptional plural oxen.

Another difficulty with rule approaches is that without additional interpretations of how to use the rules, we cannot make predictions about behavior when a given context is deviant in some way or when crucial contextual information is missing. Yet there is abundant evidence from language usage that we can interpret improperly-formed contexts, such as slips of the tongue and ungrammatical sentences, including most of the starred sentences constructed by linguists. In addition, we can usually understand non-native and dialectal speech, proving these speech types are reasonably close to our own.

We can also use redundant information to predict outcomes when normally expected information is missing. Consider the following partially-obscured word from McClelland and Rumelhart:[7]

\[
\text{WORK}
\]

Even though the final letter is partially covered, it is obvious from what is given that the word must be work, not word nor the impossible worr. Similarly, we can usually delete all the vowels from sentences of English without totally impairing our ability to understand what is intended, as in the following passage from Chomsky:[8]

\[
\text{The problem for the linguist, as well as for the child learning the language, is to determine from the data of performance the underlying system of rules that has been mastered by the speaker-hearer and that he puts to use in actual performance.}
\]

Traditionally, linguistic analyses have been based on the idea that language is a system of rules. Saussure, of course, is well-known as an early proponent of linguistic structuralism. Yet linguistic structuralism did not originate with Saussure -- nor did it end with "American structuralism". The neogrammarian approach to historical change is clearly structuralist. And it must be recognized that Chomsky himself is a structuralist par excellence. His attack against the American structuralists was not an attack against structuralism per se, but instead was an attack against the methodological assumptions that these linguists had espoused. For Chomsky (and virtually all other linguists today) there is no doubt that language is rule-governed
and that language behavior must be accounted for in terms of explicit rules. As a corollary, language acquisition is viewed as learning rules and language change as a change in the rules.

Within the structuralist framework, linguists have usually operated under the hypothesis that language rules are deterministic. Cases of variation have long been recognized, but until fairly recently these cases were vaguely identified as "free variation", a term which essentially meant that the behavior was non-deterministic. Labov and his co-workers have provided many examples of language variation which cannot be reduced to deterministic explanation. The question no longer is whether there is probabilistic behavior in language. Instead the question is: How do we account for this behavior? Following the traditional assumption that language is rule-governed, Labov has proposed that variable rules be used to account for probabilistic behavior. Yet there are some serious conceptual difficulties with variable rules.

As I see it, there are two specific problems that have arisen in the study of language variation. One is the assumption that multiple-factor effects are simple mathematical functions of single factors. This is, of course, a very specific assumption and one wonders whether there is much empirical evidence from language learning for such an underlying separation of variables. This assumption also explains the lack of reference to standard statistical approaches (such as discrete multivariate analysis) which directly consider the possibility of multiple effects.

The second specific problem follows from the first; namely, the question of which model (additive, multiplicative non-application, or multiplicative application) fits the data best -- or, equivalently, which parameters should be assigned to a logistic model. There appears to be little evidence for a principled basis on which to choose the appropriate model.

But there are also two general problems that have arisen in variation theory. One is the status of probabilities in the model -- do they actually exist? and if so, how are they learned from the statistics and then used to predict behavior? Many have argued that the probabilities do not in fact exist, but this still leaves us with the question of what does account for the probabilistic behavior.

But probably the most serious problem is the seemingly unlimited number of variables (linguistic as well as social) that affect any given non-deterministic phenomenon in language. In other words, there seems to be no end to the variation. This reminds one of the well-known suspicion in statistics that completely independent variables are rare, that given enough data almost any two variables can be shown to be dependent to some degree.
As a specific example of this problem, consider Guy's discussion of final /t,d/ deletion in English. Guy first lists Labov's three main linguistic variables that affect the probability of final /t,d/ deletion. Originally, Labov assigned two variants to two of these linguistic variables and ignored the lesser effect of the third:

1. **Grammatical conditioning**: Is the word monomorphemic or bimorphemic?
2. **Conditioning by following segment**: Is the segment a vowel or a non-vowel?
3. **Conditioning by preceding segment** (not specified in Labov's original study)

Over the years additional studies have shown that each of these variables should be assigned more variants — that additional distinctions are necessary if we want to account for empirically significant effects on final /t,d/ deletion. Guy partitions these three variables more finely:

1. **Grammatical conditioning**: Is the word monomorphemic, the past tense of an ambiguous verb, or the past tense of a regular weak verb?
2. **Conditioning by following segment**: Is the segment a consonant, a glide, a liquid, a vowel, or a pause?
3. **Conditioning by preceding segment**: Is the segment a sibilant, a non-sibilant fricative, a nasal, a stop, or a lateral?

So from an original 4-way distinction we now have a potential distribution of 75 possibilities ($3^3 = 75$).

But this is not all. Other variables have been discovered that affect the probability of final /t,d/ deletion: lexical stress, rate of speech, length of consonant cluster, articulatory complexity of clusters, speech style, and social factors (such as age, sex, social class, race, geographic background, and so on).

Guy rightly observes that this rule of final /t,d/ deletion shows that "variation is inherent, and cannot be scrubbed out of our linguistic description by ever-finer subdivisions of the data." In other words, final /t,d/ deletion cannot be reduced to deterministic phenomena. But the history of this example also implies that final /t,d/ deletion cannot be correctly described either! There doesn't seem to be any limit to the number of variants that affect final /t,d/ deletion. As more data is collected, more distinctions are discovered. The effect of these additional variants is less important, but they are
still statistically significant. It appears as if no simple correct description of final /t,d/ deletion is forthcoming. In attempting to describe non-deterministic phenomena optimally, we will have to sacrifice minimality in order to achieve correctness.

This example seems to point to the following conclusion -- that ultimately the correct description may have a separate rule for every different set of conditions. Taken to its logical conclusion, this would mean that each rule would represent a single occurrence since no two occurrences are completely identical. In other words, instead of representing types of occurrence, rules would represent tokens of occurrence.

In the second part of this paper this idea will be developed as an alternative to structuralist descriptions. In fact, the notion of rule will be abandoned in favor of an analogical approach that avoids the conceptual difficulties of rule approaches. Instead of trying to predict behavior by using a system of rules (probabilistic or otherwise), an analogical description predicts behavior by means of a collection of examples called the analogical set. Given a context \( x \), we construct the analogical set for \( x \) by looking through the data for

1. classes of examples that are most similar to \( x \), and
2. more general classes of examples which behave like those examples most similar to \( x \).

In order to show how to construct the analogical set for a given context, let us consider final stop deletion once more. Suppose we are interested in predicting final stop deletion when we are given the context \( \bar{v}w\bar{s} \) -- that is, when the final stop is not followed by a vowel (\( \bar{v} \)), is not word-final (\( \bar{w} \)), and is preceded by a sonorant (\( s \)). For this given context, we construct a hierarchy of supracontexts by systematically eliminating the factors \( \bar{v}, \bar{w}, \) and \( s \). For each of these supracontexts we also determine the number of times the final stop is deleted and the number of times it is retained:

\[
\begin{array}{c}
\bar{v}w\bar{s} \quad (21,4) \\
\bar{v}\bar{w} \quad (34,6) \quad \bar{v}-s \quad (58,14) \quad \bar{w}s \quad (28,9) \\
\bar{v} \quad (117,37) \quad \bar{w} \quad (55,16) \quad --s \quad (80,67) \\
--- \quad (172,198)
\end{array}
\]
We next test each of these supracontexts for homogeneity. A supracontext is homogeneous providing all its subcontexts (as defined by the given context) behave alike. Using a statistical procedure that I will explain momentarily, we obtain the following results for this hierarchy of supracontexts:

(1) The given context (\(\bar{\text{vws}}\)) is by definition homogeneous since it has no subcontexts.

(2) There are two supracontexts that are statistically homogeneous: \(\bar{\text{vws}}\) and \(\bar{\text{v}}\)-s.

(3) Two supracontexts are statistically heterogeneous: \(\bar{\text{w}}\)-s and \(\bar{\text{v}}\)-s. These supracontexts have been circled.

(4) If a supracontext is statistically heterogeneous, then all supracontexts containing that supracontext are automatically considered heterogeneous. In our hierarchy there are three such inclusively heterogeneous supracontexts: \(\bar{\text{w}}\)-s, \(\bar{\text{v}}\)-s, and \(\bar{\text{v}}\)-s. These supracontexts are enclosed by boxes.

Now the analogical set is formed by all the occurrences from each homogeneous supracontext. In our example we obtain 137 occurrences, of which 113 involve final stop deletion:

\[
\begin{array}{ccc}
\text{\(\bar{\text{vws}}\)} & 21 & 4 \\
\text{\(\bar{\text{v}}\)-s} & 34 & 6 \\
\bar{\text{v}}\)-s & 58 & 14 \\
\hline
113 & 24
\end{array}
\]

In order to predict behavior, we posit a rule of usage called random selection:

Randomly select one occurrence from the analogical set and use it as a model for predicting behavior.

With this rule of usage, we obtain the probability that the final stop will be deleted when the given context is \(\bar{\text{vws}}\):

\[
P(\text{deletion} \mid \bar{\text{vws}}) = 113/137 \approx 0.825
\]

But the question still remains: How do we statistically determine whether a supracontext and its subcontexts behave alike. Traditionally, in using a statistical test, we must know either the underlying probability distribution for the test or a distribution that approximates the actual distribution. In this paper, I use a natural statistic called the rate of agreement.
that avoids this problem. If we have a context with \( n_i \) occurrences of each outcome \( \omega_i \), with \( n = \sum n_i \) occurrences in all, then the rate of agreement is defined as

\[
\ell = \frac{1}{n-1} \sum n_i (n_i-1)/2
\]

Using this statistic, our decision procedure for testing homogeneity of behavior turns out to be very simple:

Always try to increase the rate of agreement.

To show how this works, consider the following array of data derived from our hierarchy of supracontexts:

<table>
<thead>
<tr>
<th></th>
<th>21</th>
<th>4</th>
<th>25</th>
</tr>
</thead>
<tbody>
<tr>
<td>vws</td>
<td>7</td>
<td>5</td>
<td>12</td>
</tr>
<tr>
<td>-ws</td>
<td>28</td>
<td>9</td>
<td>37</td>
</tr>
</tbody>
</table>

For each context we calculate the rate of agreement:

- \( \ell (vws) = (1/24) [21 \cdot 20 + 4 \cdot 3]/2 = 9 \)
- \( \ell (vws) = (1/11) [7 \cdot 6 + 5 \cdot 4]/2 = 2.818 \)
- \( \ell (\bar{-}ws) = (1/36) [28 \cdot 27 + 9 \cdot 8]/2 = 11.5 \)

Since the two subcontexts (\( vws \) and \( v\bar{w}s \)) form a partition on the supracontext (\( \bar{-}ws \)), we sum up the rates of agreement for the two subcontexts and compare that sum with the rate of agreement for the supracontext:

\[
\Delta \ell = \ell (vws) + \ell (v\bar{w}s) - \ell (\bar{-}ws)
\]

\[
\approx 9 + 2.818 - 11.5
\]

\[
\approx 0.318
\]

Since we always try to increase the rate of agreement, we have the following decision procedure:

- If \( \Delta \ell > 0 \), then the array is statistically heterogeneous.
- If \( \Delta \ell < 0 \), then the array is statistically homogeneous.

Therefore, our sample array is statistically heterogeneous.

This statistical procedure is very simple. It never requires us to calculate probabilities or use an approximate distribution to estimate those probabilities. Another advantage
is that it is parameter-free. Its level of significance is asymptotically less than one half, but nonetheless fairly close to one half. This of course means that this test is very powerful. From a decision point of view this procedure can be defended in that it equally favors heterogeneity and homogeneity -- unlike traditional statistical procedures which strongly favor homogeneity. (It also turns out that we can redefine this statistical test so that decisions are made at smaller levels of significance.)

This procedure is also biased towards deterministic predictions of behavior, especially under certain well-defined conditions; namely,

(1) when the number of occurrences is low; or

(2) when imperfect memory reduces the number of occurrences.

This biasedness towards deterministic predictions helps explain several well-known observations about language variation; for example,

(1) the historical tendency to replace synchronic language variation by deterministic behavior; and

(2) the historical tendency to split up a fairly frequent non-deterministic context into a class of deterministic contexts.

In addition, deterministic behavior is favored whenever there is a need to maximize utility (that is, maximize rewards or minimize losses). Under such conditions a different rule of usage called selection by plurality is used:

Select the outcome which occurs most frequently in the analogical set and predict that the given context will take that outcome.

In order to compare this analogical approach to a rule approach, let us consider an example of morphological variation from Finnish. In Finnish certain bisyllabic verb stems ending in

[non-obstruent] [dental stop] [low vowel]

take imperfect forms ending in ti or si. Some of these verbs take only ti, some only si, and others can take either ending but with different degrees of acceptability (depending on the particular verb). The contextual space for this class of verbs can be split up into four distinct categories according to the length of the vocalic portion (either short V or long VV) and whether or not the syllabic ends in a consonant C:
The last category, VVC, is statistically deterministic since it only takes the \( sj \) outcome (in the standard language). An optimal structuralist description of this class of verbs would at least combine all the verbs in the VVC category into a single rule. Such a rule would predict that a verb like \( \text{viertaa} \), would take the ending \( sj \) in the imperfect. In contrast to the rule approach, consider the analogical set for \( \text{viertaa} \) (based on statistics from Tuomi's statistical analysis of standard Finnish):[16]

<table>
<thead>
<tr>
<th>distance from verb</th>
<th>verb</th>
<th>number of supracontexts</th>
<th>frequencies in analogical set</th>
<th>probability of random selection</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>ti</td>
<td>si</td>
<td>ti</td>
</tr>
<tr>
<td>0</td>
<td>( \text{viertaa} )</td>
<td>0</td>
<td>0</td>
<td>38</td>
</tr>
<tr>
<td>1</td>
<td>( \text{kiertaa} )</td>
<td>0</td>
<td>16</td>
<td>22</td>
</tr>
<tr>
<td>2</td>
<td>( \text{piirtaa} )</td>
<td>0</td>
<td>24</td>
<td>14</td>
</tr>
<tr>
<td>3</td>
<td>( \text{siirtaa} )</td>
<td>0</td>
<td>8</td>
<td>14</td>
</tr>
<tr>
<td>4</td>
<td>( \text{kieltaa} )</td>
<td>0</td>
<td>22</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>( \text{viiltaa} )</td>
<td>0</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>6</td>
<td>( \text{pyoirtaa} )</td>
<td>0</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>( \text{kiiltaa} )</td>
<td>0</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>( \text{vaanta} )</td>
<td>0</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>( \text{murtaa} )</td>
<td>0</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>( \text{sortaa} )</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>( \text{huoltaa} )</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>12</td>
<td>( \text{juontaa} )</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>13</td>
<td>( \text{kuultaa} )</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>14</td>
<td>( \text{kaanta} )</td>
<td>0</td>
<td>26</td>
<td>2</td>
</tr>
<tr>
<td>15</td>
<td>( \text{muuntaa} )</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>16</td>
<td>( \text{myonta} )</td>
<td>0</td>
<td>30</td>
<td>2</td>
</tr>
<tr>
<td>17</td>
<td>( \text{puoltaa} )</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>18</td>
<td>( \text{tyonta} )</td>
<td>0</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>19</td>
<td>( \text{santa} )</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

\[ P(\text{silviertaa}) \approx .9985 \]

We first observe that the predicted behavior for \( \text{viertaa} \) looks rule-governed. In the analogical approach the predicted probability for the \( sj \) outcome is very close to one, the value that an optimal rule description of the data would predict.

But we also note that the analogical approach is, in a sense, messy: it permits the verbs \( \text{murtaa} \) and \( \text{sortaa} \) (which do not have the same syllabic structure as \( \text{viertaa} \)) to affect the predicted outcome. In contrast to an optimal rule approach, the
The analogical description allows a slight "leakage" in favor of the \( \ddagger \) outcome.

The analogical approach also assigns probabilities to particular proportional analogies. The preferred analogy is, in the case of \textit{viertäät}, a rhyming analogy:

\[
\text{kierätä : kiersi :: viertäät : viersi}
\]

This analogy occurs 27\% of the time. Rhyming analogies tend to be fairly significant, especially when the given context is non-occurring. But we must keep in mind that the analogical approach does not necessarily prefer rhyming analogies: the occurrence of such analogies depends on whether rhyming contexts are homogeneous in behavior.

This example also shows that the effect of a particular verb depends upon three factors:

(1) the amount of similarity between the verb and the given context;

(2) the frequency of the verb; and

(3) the number of homogeneous supracontexts that the verb occurs in (or, equivalently, the extensiveness of the homogeneity).

These same multiple analogical effects were noticed by Johnson and Venezky in their study of speakers' pronunciation of unfamiliar words:[17]

A model that might provide a higher degree of predictability ... is a final consonant model [in essence, a rhyming model] based on token counts rather than on type counts. This model would be especially effective if the final consonant influence derives from analogy with a few high frequency words rather than from a generalization based on all real words that contain a particular spelling.

This model is, of course, compatible with an analogical approach based on frequency of occurrence.

In many cases (like this one) the predicted behavior is nearly the same no matter which approach is used, but conceptually the two approaches are quite different:
**RULE APPROACH**

- A system of rules
- Based on types of behavior
- Contextual space is partitioned
- Global, macroscopic
- Must learn rules from the data
- Static, rigid
- Usage: find the applicable rule
- Must specify how rules interact
- Well-defined boundaries
- Sharp, precise transitions
- Rule-governed
- Predictions made by rules alone
- Explicit, direct

**ANALOGICAL APPROACH**

- Equivalent to the original data
- Based on tokens of behavior
- Contextual space remains atomistic
- Local, microscopic
- Need for large memory capacity
- Dynamic, flexible
- Usage: find an appropriate example
- Must be able to access data quickly
- No boundaries directly defined
- Gradual, fuzzy transitions
- Appears to be rule-governed
- Predictions only for given contexts
- Implicit, indirect

Many of these same distinctions are found in Winograd's terms "declarative" versus "procedural".[18] This distinction seems to be particularly relevant in explaining language performance, as has been pointed out by Rumelhart:[19]

Perhaps the classical case of using knowledge how (procedural knowledge) to produce knowledge that (factual knowledge) occurs in the domain of grammatical judgements. The knowledge that we have about language seems to be largely embedded in the procedures involved in the production and comprehension of linguistic utterances. This is evidenced by the relative ease with which we perform these tasks when compared with our ability to explicate the knowledge involved in them. Semantic knowledge would appear to be the same. Whereas we can quickly interpret sentences, it is only with the most painstaking effort that we can produce definitions of terms with any generality.

Despite these arguments, both empirical and conceptual, in favor of an analogical approach to the description of language (as well as other forms of behavior), there is a place for structuralism too. Structuralist descriptions are properly used to describe actually-occurring behavior. An optimal structuralist description serves as a kind of meta-language that efficiently describes past behavior and allows us to talk about that behavior. Whenever we attempt to summarize behavior or to discover relationships in data, our viewpoint is structuralist. But if we wish to predict behavior rather than just describe it, it may be necessary to abandon structuralist approaches in favor of an atomistic one.
NOTES

1 This paper is based on my forthcoming book *Analogy and Structure*.

2 Abramson and Lisker 1972:19.


4 Gates 1937:72.

5 I owe this example to material provided by Thomas D. Horn.

6 Johnson and Venezky 1976:262.


9 Labov 1969:737-739.

10 Cedergren and Sankoff 1974:335-336.


14 Guy 1980:11.


16 Tuomo Tuomi, statistical analysis of approximately 600,000 words of text from 1975-76 issues of *Suomen Kuvalehti* (available on microfiche).

17 Johnson and Venezky 1976:266.


REFERENCES


