Analogy: A non-rule alternative to neural networks

Royal Skousen

This paper provides a general introduction to analogical modeling of language, first comparing it with rule approaches, the traditional method of language description. An alternative procedural approach to language description is found in neural networks, but this approach contains a number of serious design defects. The current status of research on analogical or exemplar-based modeling is also presented.[1]

1. Introduction

In this paper I would like to discuss three aspects of analogical modeling. First, I will provide a basic introduction to the analogical approach, comparing it with rule approaches. I will then compare analogy with neural networks (also known as connectionism). Although both these approaches are procedural alternatives to rule-based approaches, there are some significant differences between the two. Finally, I will discuss some of the current problems in developing procedural approaches.

2. Three Basic Types of Behavior

In order to contrast the basic differences between the rule approach and the analogical one, we will consider how each approach deals with three basic kinds of behavior:

(1) *categorical* (such as the indefinite article *a / an* in English);

(2) *exceptional / regular* (such as the spelling of the word-initial /h/ sound in English);

(3) *idiosyncratic* (such as the voicing onset time – or simply VOT – for /b/ versus /p/ in English).

In the rule approach, there is a demarcation of boundaries such that the contextual space is divided up (or partitioned) into a set of explicit rule contexts, as in each of the three basic types of behavior:

Rivista di Linguistica, 7.2 (1995), pp. 213-231 *(ricevuto nell'ottobre 1994)*

(1) *categorical*: the selection of *a* or *an* depends on whether the initial segment of the following word is a consonant or a vowel;

(2) *exceptional / regular*: the spelling of the word-initial /h/ sound gives a general case (namely, the spelling <h>), plus a list of exceptions (*who, whole, whore, whooping, Jose,* ...);

(3) *idiosyncratic*: the phonemic difference between /p/ and /b/ is determined by listing certain regions of voicing onset time (in milliseconds) for each phoneme, but leaving other regions undefined: /b/ [-130,0], /p/ [20,90].

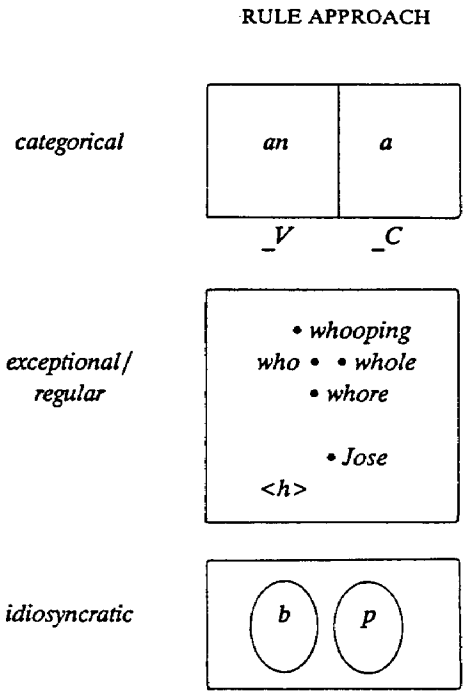

Figure 1.

In order to use a rule to predict behavior for any given case, we find the appropriate rule context and then apply the rule. In the case of the indefinite article in English, we look for the first segment of the following word and predict *a* or *an* according to whether that segment is a consonant or a vowel. In the case of the spelling of word-initial /h/, we first determine if the word is in the list of exceptions (for example, *who,*

whole, whore, whooping, and *Jose*). If it is, then we spell the /h/ according to the particular exception. Otherwise, we use the regular spelling <h> to spell the /h/. In the VOT example, we determine whether the given labial stop has voicing begin in the interval [-130,0] or in the interval [20,90] and then predict /b/ or /p/ accordingly.

3. *Problems with Rules*

There are a number of behavioral problems with the rule approach. First of all, there is the problem of "leakage" across categorical boundaries. For example, in the case of the indefinite article in English, there is a tendency for *an* to be replaced by *a* (as exemplified in children's language, performance errors among adults, and dialect developments). This leakage is directional: there is little tendency for *an* to replace *a*.

Second, items close to exceptions can occasionally behave exceptionally, as in the common children's misspelling GREAD for *grade* (based on *great*), or the morphological form *axen* as the plural of *ax* (based on *ox / oxen*).

Third, speakers have the ability to predict behavior in cases of idiosyncracy where no outcome is defined, as in the VOT experiments of Lisker and Abramson, or in Labov's semantic experiments with cups and bowls.

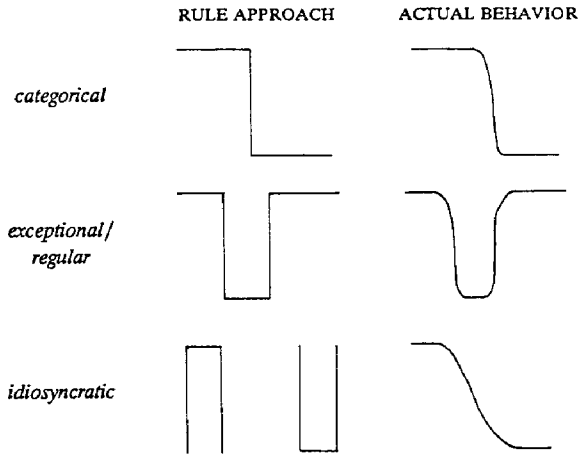

Figure 2.

(For further discussion of these examples and others like them, see Skousen 1989:15-18)

4. The Analogical Alternative

Instead of trying to determine the behavior for the contextual space, the analogical alternative stores exemplars in a multivariate contextual space and then uses these examples to predict behavior for specific given contexts. In the following diagrams, we see how the contextual space is viewed by the analogical approach:

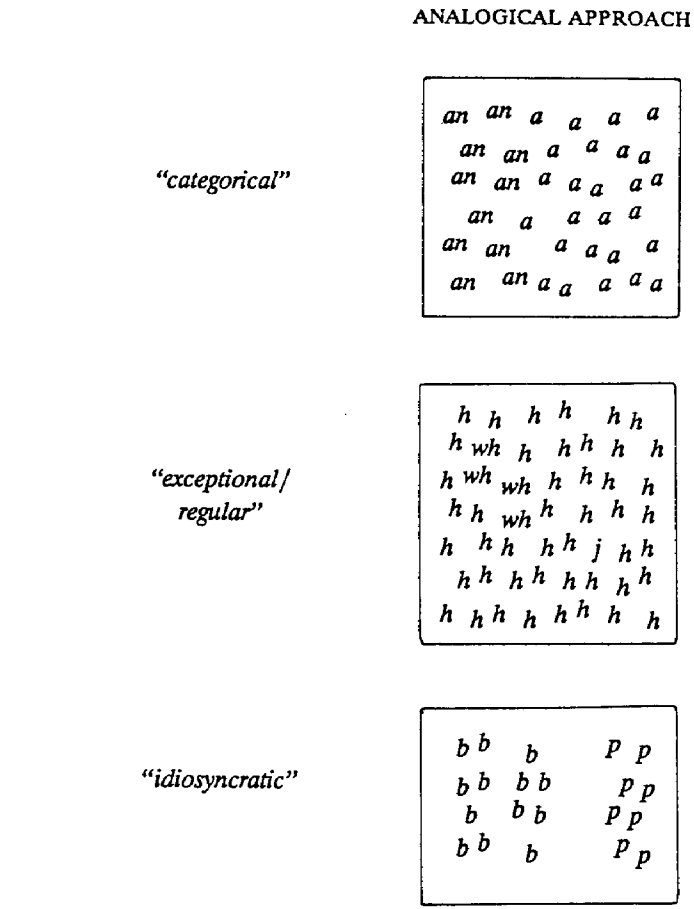

Figure 3.

It should be noted that in each of these representations a multivariate contextual space has been reduced to a two-dimensional diagram. In the

case of the *a / an* example, the placement of the occurrences of *a* and *an* has been made according to the rule-based variable (that is, according to the syllabic nature of the first segment of the following word). Thus the *a*'s and *an*'s are separated in this diagram. On the other hand, if the two-dimensional diagram were drawn according to any other variable (such as the syllabic nature of the last segment of the preceding word), then the *a*'s and *an*'s would generally be randomly mixed throughout the two-dimensional diagram.

In using the analogical approach, we can predict behavior only for a given context. We first search for actual examples of that context and then move outward in the contextual space looking for nearby examples. In working outward away from the given context we systematically eliminate variables, thus creating more general contexts called *supracontexts*. The examples in a supracontext will be accepted as possible analogical models only if the examples in that supracontext are homogeneous in behavior.

5. Properties of Analogical Models

Three important properties affect the probability of selecting a particular example as an analogical model (see Figure 4 on the next page):

(1) *proximity*: the more similar the example is to the given context, the greater the chances of that example being selected as the analogical model;

(2) *gang effect*: if the example is surrounded by other examples having the same behavior, then the probability of selecting these similarly behaving examples is substantially increased;

(3) *heterogeneity*: an example cannot be selected as the analogical model if there are intervening examples, with different behavior, closer to the given context.

The property of heterogeneity is very important. Traditional appeals to analogy in linguistic description have, in theory at least, permitted any example to serve as an analogical model. But in this explicit approach to analogy we determine the probability of using any particular example. And in certain cases, because of heterogeneity, some examples will be totally excluded from the set of analogical models for a particular given context.

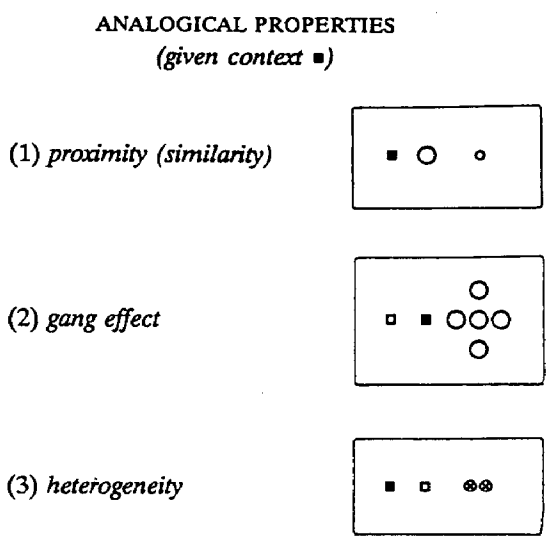

Figure 4.

6. Natural Statistics

It turns out that all three of the above analogical properties are derived from a very simple principle of minimizing disagreement. The traditional measure of uncertainty is the logarithmic measure of entropy (also known as Shannon's information). This measure is based on the idea that one gets an unlimited number of chances to guess the correct outcome. I reject this measure because of its psychological implausibility. In my analogical work I propose a very simple quadratic measure of disagreement which is based on the idea that one gets a single chance to guess the correct outcome. Heterogeneity in the contextual space occurs whenever there is an increase in the number of disagreements. The two other properties (of proximity and the gang effect) are also derived from this measure of disagreement.

It turns out that this decision procedure (of minimizing the number of disagreements) represents the most powerful statistical test possible. Any supracontext that could possibly be heterogeneous is declared to be heterogeneous. In fact, this decision procedure eliminates any need for statistical calculation at all. Of course, this approach is completely contrary to all standard statistical procedure. But by adding imperfect memory to the data set, the analogical procedure becomes equivalent to standard statistics. In particular, a statistically significant relationship is one that is still predicted even when memory is poor (that is, when

most of the data is forgotten or ignored). I call this kind of statistical procedure a "natural statistic" since it avoids all the complications of standard statistics, yet behaves in accord with those standard statistics. Such natural statistics are psychologically plausible and therefore suitable for a linguistic theory that attempts to predict actual language behavior (for further discussion of the statistical basis for analogical modeling, see chapters 2 and 4 in Skousen 1989, as well as chapters 11-13 and 16 in Skousen 1992).

7. A Prototypical Categorical Rule

We will now briefly consider some applications of the analogical approach to various language problems. First, let us consider the analogical treatment of a prototypical rule, such as the indefinite article *a/an* in English (in this example, I follow the discussion in section 3.2 of Skousen 1989). We first construct a data sample based on actual examples. In this case the data set is derived from a paper I wrote in 1985 in which there are 136 examples of *a* and 28 of *an* (for a total of 164). For each example, I specify nine variables. These variables basically represent phonological characteristics in proximity to the indefinite article. Distinctive features are not represented, but phonemes and syllabic structure are (for discussion of how variables are selected, see section 3.1 of Skousen 1989). The nine variables describe the two preceding segments and the two following ones, and also specify the type of syntactic transition just prior to the indefinite article:

> *variables 1 and 3:* the syllabic category (Consonant or Vowel) of the two immediately preceding phonemes;
>
> *variables 2 and 4:* the two immediately preceding phonemes;
>
> *variable 5:* whether the indefinite article begins an independent phrase (|) or not (+);
>
> *variables 6 and 8:* the syllabic category (Consonant or Vowel) of the two immediately following phonemes;
>
> *variables 7 and 9:* the two immediately following phonemes.

Thus the context *through a glass darkly* has the outcome *a* and the nine variables CrVu+CgCl; similarly, the context *night — an inn* has the outcome *an* and the nine variables VaiCt|ViCn. In all, the data set contains 164 such specifications.

On the basis of this data set of 164 examples, the analogical

approach easily makes predictions for various given contexts. For example, for the given context *picked ? apple* (which has the nine-variable specification CkCt+VæCp), the predicted outcome is *an* 99.2% of the time and *a* 0.8% of the time. In other words, the analogical approach basically predicts the rule-based outcome *an*, but with some slight leakage towards *a*. On the other hand, for the given context *picked ? pear* (CkCt+CpVɛ), the predicted outcome is only *a*, with no leakage at all towards *an*. This generally holds for other given contexts as well. If the following word begins with a vowel, we tend to get *an* about 98-99% of the time, with a 1-2% leakage towards *a*; but if the following word begins with a consonant, we virtually always get *a*, with no leakage towards *an*.

This same instability of the *an* form of the indefinite article also shows up when we consider this data set from an acquisitional point of view. If the data set contains less than 70 randomly selected examples of the indefinite article, there is a good chance of getting *a* rather than *an* when the following word begins with a vowel. Moreover, for such small numbers of occurrences, there is a high degree of variability and not much patterning. But for more than 70 examples, a certain level of stability is achieved. Leakage of *an* remains, but at a fairly constant rate of 1-2%. Nor does there seem to be any further improvement in reducing this leakage as the number of occurrences increases. On the other hand, when the following word begins with a consonant, we get *a* virtually 100% of the time, even when the number of examples in the data set is small. No matter what the size of the data set, the leakage is directional (from *an* to *a*, with hardly ever *a* to *an*).

8. *The Competence-Performance Distinction*

Now if we wish to "explain" this non-symmetric behavior from a rule perspective, we must move beyond the rules themselves since the rules are symmetric in form: *a* / __ *C* versus *an* / __ *V*. In order to predict behavioral aspects, we need to add on some performance factors that will prefer the *a* outcome over the *an* outcome. We could establish some kind of formal or taxonomic principles (such as "markedness" or conditions on rule formulation) that might prefer, say, open syllable forms (such as *a*) over closed syllable ones (such as *an*). But the rules themselves are inherently incapable of predicting the actual non-symmetric behavior. In fact, the rules are often declared to represent some kind of abstract (or non-empirical) "competence". The rules per se do not have any built-in empirical interpretation, so the performative aspects must be added. In actuality, the whole competence-performance distinction may well be

the result of first trying to use rules to describe language, and then adding performative factors to patch up the failure of the rules to account for dynamic aspects of language (such as drift, fuzzy boundaries, leakage, and so forth).

On the other hand, the analogical approach makes no distinction at all between competence and performance. The same analogical procedure produces the rule-governed behavior as well as the tendencies towards change (including historical drift, dialect development, adult errors, and children's language).

9. *Robustness*

One particular aspect where the rule system fails is when the crucial context is missing. Speakers have the ability to predict an outcome even when the given context is missing information or is in some sense ill-formed. Suppose in our *a* / *an* example that we wish to predict the indefinite article, but for some reason the first segment of the following word is obscured so that we cannot tell whether it is a consonant or a vowel. In the rule approach, no prediction is possible since the crucial contextual specification is missing. Without adding a plethora of idiosyncratic rules to handle cases when the crucial variable is missing, the rule approach will totally fail. But the analogical approach is perfectly capable of making a prediction when the "crucial" variable is missing. On the basis of the remaining variables, it still constructs an analogical set; and in most instances it makes the prediction that would be correct if the initial segment were not obscured. This success of the analogical approach essentially derives from its ability to recognize the following word even when the first segment is obscured. Occasionally, it is also able to take advantage of phonological tendencies in the language (such as expected syllabic structure). In any event, the analogical approach does not collapse when confronted with an ill-formed or incomplete given context.

This issue of robustness is a very serious one. Rule approaches are just not robust. Speakers clearly demonstrate an ability to deal with language when information is missing or when the given context is ill-formed. Words can be recognized even when parts of letters or even whole letters are missing. The vowels in an English language text can be replaced by asterisks, yet the text remains readable for the most part. Speakers can understand slips of the tongue and, in most cases, the starred sentences of generative syntacticians (for specific examples, see section 1.3 of Skousen 1989).

10. Probabilistic Behavior

Suppose we have a given context that occurs and the behavior for this context is probabilistic (that is, there are examples of more than one outcome). Many linguistic examples of probabilistic behavior have been identified (especially by William Labov and his colleagues). There is no question as to the existence of non-deterministic linguistic behavior. The problem is how to deal with such linguistic variation. From a rule perspective, we need to figure out how a probability can be learned and then used by speakers to predict occurrences probabilistically (for further discussion of this problem, see sections 4.1 and 4.4 in Skousen 1989).

From the analogical point of view, the solution is to avoid directly learning and using probabilities. Rather, we store examples and then randomly select one of these examples to predict the outcome. Now if we have perfect memory, then the probability of predicting a specific outcome will be exactly equal to its relative frequency in the set of analogical examples. But if we have imperfect memory, then our chances of predicting a specific outcome will vary from its relative frequency, but within certain probabilistic limits (which turn out to be asymptotically equivalent to standard statistical variability when the probability of remembering any particular occurrence is one-half). (See section 4.3 of Skousen 1989 for the effect of imperfect memory on probabilistic prediction).

11. Multivariate Analysis of Linguistic Variation

Actually the more difficult question is how to make probabilistic predictions when the given context does *not* occur. This situation often occurs in cases of complex sociolinguistic variation. In section 4.5 of Skousen (1989) I discuss such a case of variation in colloquial Cairo Arabic: namely, the probabilistic behavior of two terms of address, *ya'axi* and *yaxuuya*, which both mean 'my brother'. These two terms are in competition with each other, but the probability of selecting *ya'axi* or *yaxuuya* depends on a number of social variables (the sex, social class, and age of the speaker, the social class and age of the addressee, the speaker's tone of voice, the power relationship defined by the social situation, and the degree of closeness or familiarity between the speaker and the addressee). As is typical of linguistic variation, each occurring combination of social variables is probabilistic and cannot be reduced to deterministic description. Moreover, given enough data, each occurring combination seems to take a different probability. And, of course, many

of the potential combinations of social variables are not found in the data set.

Now the question is: What is the relationship between the many sociolinguistic variables that seem to affect the outcome? One approach could be to directly assign a separate probability for every possible combination of variables. This solution seems inordinately unwieldy to variationists working within the rule framework. Moreover, such a solution fails to define a probability whenever we have a non-occurring combination of variables. So the only apparent alternative has been to assume that each single variable has a probability of occurrence assigned to it, and then some general mathematical relationship is used to calculate the probability of occurrence for groups of variables from the probabilities of the single variables. This approach forms the basis of David Sankoff's VARBRUL program, which proposes that the general mathematical relationship is a logistic one with one freely varying parameter that is manipulated to get the best possible agreement between predicted behavior and actual behavior. Of course, there is really no empirical evidence for this whole procedure; and its psychological plausibility is quite dubious.

Although specific probabilities are not learned, the analogical approach otherwise does what the variationists have avoided doing: namely, specifying outcomes for fully-specified combinations of variables. In the Arabic example, I used 242 examples (collected by Dil Parkinson) of the two terms for 'my brother'. For each of these 242 occurrences, the nine social variables listed above were specified. On the basis of this data set, I was then able to make analogical predictions for *ya'axi* and *yaxuuya* for both occurring and non-occurring combinations of variables – and without ever have to directly discover the "relationships" between the variables. Moreover, statistical significant predictions held under conditions of very poor memory, whereas random insignificant predictions vanished under relatively poor memory.

12. Predicting Unexplained Behavior

Analogical models also have the capability of correctly predicting behavior that appears to be anomalous or unexpected from a rule approach, especially when that approach identifies certain variables as being "crucial" while other variables are ignored as "insignificant". A good example of this problem occurs in the Finnish past tense. If a two-syllable verb stem ends in a non-high front unrounded vowel preceded by a consonant (namely, *Ce, Cä,* or *Ca*), three different past-tense forms are possible, providing certain additional conditions are met:

a-oi: if the final vowel is *a*, the *a* can be replaced by *oi* in the past tense, especially when the first vowel is *a* (as in *laula-* 'sing', with *lauloi* as the past tense);

tV-si: if the final consonant is *t* and is preceded by a non-obstruent, the *t* and the final vowel can be replaced by *si* in the past tense (as in *lentä-* 'fly', with *lensi* as the past tense);

V-i: or, in general, the final vowel can be replaced by *i* (as in *luke-* 'read', with *luki* as the past tense).

The problem with using rules to describe the past tense in Finnish is that such rules are often incapable of predicting what actually occurs, especially for infrequent verbs in the language. But the analogical approach readily makes the right predictions – and can even tell us why one particular past tense form is preferred over another. In the analogical approach, we predict behavior on the basis of the frequent verbs in the language (in this case, 173 of them), with enough variable specification to account for the phonemic and syllabic structure of the complete verb stem (see section 5.4 in Skousen 1989 for how to construct the Finnish data set).

Consider a case from Finnish where the rule approach is unable to distinguish between two types of verbs. There are three relatively infrequent verb stems (not in the list of 173) that have a long *a* as the first vowel and end in *ta* preceded by a non-obstruent: *kaarta* 'swerve', *saarta-* 'surround', and *raata-* 'toil'. In the case of the first two verbs, the analogical approach predicts that both the outcomes *a-oi* and *tV-si* are about equally possible, but in the case of the last verb, only *a-oi* is really possible:

	V-i	*a-oi*	tV-si
kaarta- 'swerve'	0	0.486	0.514
saarta- 'surround'	0.001	0.419	0.579
raata- 'toil'	0	0.996	0.004

These results agree with actual usage. For instance, the standard unabridged Finnish dictionary (Sadeniemi 1973) lists both *kaartoi* and *kaarsi* as equally possible past tense forms for *kaarta-*, and similarly *saartoi* and *saarsi* for *saarta-*. On the other hand, only *raatoi* is listed for *raata-*. Rule approaches have been written in terms of the more frequent verbs of the language, and these rules have been unable to predict why *raata-* takes only *raatoi* in the past tense – and not *raati* or *raasi*.

When we look at the analogical sets for these three verbs, we see why the analogical approach makes the distinction between *raata-* on

the one hand, and *kaarta-* and *saarta-* on the other. In the case of *kaarta-* and *saarta-* there are, in each example, two similar competing verbs that ultimately lead to a near split between the two outcomes *a-oi* and *tV-si*:

kaarta- 'swerve'	*kaata-* 'overturn'	*a-oi*
	kiertä- 'wind'	*tV-si*
saarta- 'surround'	*saatta-* 'accompany'	*a-oi*
	siirtä- 'move'	*tV-si*

But in the case of *raata-*, all of the nearby verbs (*raasta-* 'grate', *kaata-* 'overturn', *haasta* 'summon', *paahta-* 'scorch', and *saatta-* 'accompany') take the *a-oi* outcome.

Sometimes "unimportant" variables may play a significant role. Consider the relatively infrequent verb *sorta-* 'oppress' (also not in the list of 173 frequent verbs). According to all rule approaches that have been formulated, this verb should take the outcome *tV-si* in the past tense. Yet speakers favor the *V-i* outcome. Thus the unabridged Finnish dictionary (Sadeniemi 1973) lists only *sorti* as the past tense for *sorta*, not *sorsi*. Looking at our list of 173 verbs, we might think that the past tense should be *sorsi* since in the list all verb stems ending in *rtA* (that is *rta* or *rtä*) take only the *tV-si* outcome: *kiertä-* 'wind', *murta-* 'break', *piirtä-* 'draw', *pyörtä-* 'turn back', and *siirtä-* 'move'. Since *sorta-* actually takes *V-i*, how can we explain its exceptionality, especially since it is a relatively infrequent verb?

The analogical set for *sorta-* shows a surprising result. First of all, the correct outcome *sorti* is clearly predicted:

	V-i	*a-oi*	*tV-si*
sorta- 'oppress'	0.941	–	0.059

Although all five of the *rtA* verbs show up in the analogical set, they are overwhelmed by a competing (and much more massive) gang of verbs that all have *o* as the first vowel (as in figure 5 on the next page). It turns out that there are 24 verbs in the analogical set whose first vowel is o (*poltta-* 'burn', *souta-* 'row', *johta-* 'lead', and so forth) – and each of these verbs takes the *V-i* outcome! From the rule perspective, this result is purely an accident – in fact, from an historical point of view, it is an accident. Moreover, only when the analogical approach was applied to *sorta-* were we able to discover that the o vowel was causing *sorta-* to take *sorti*.

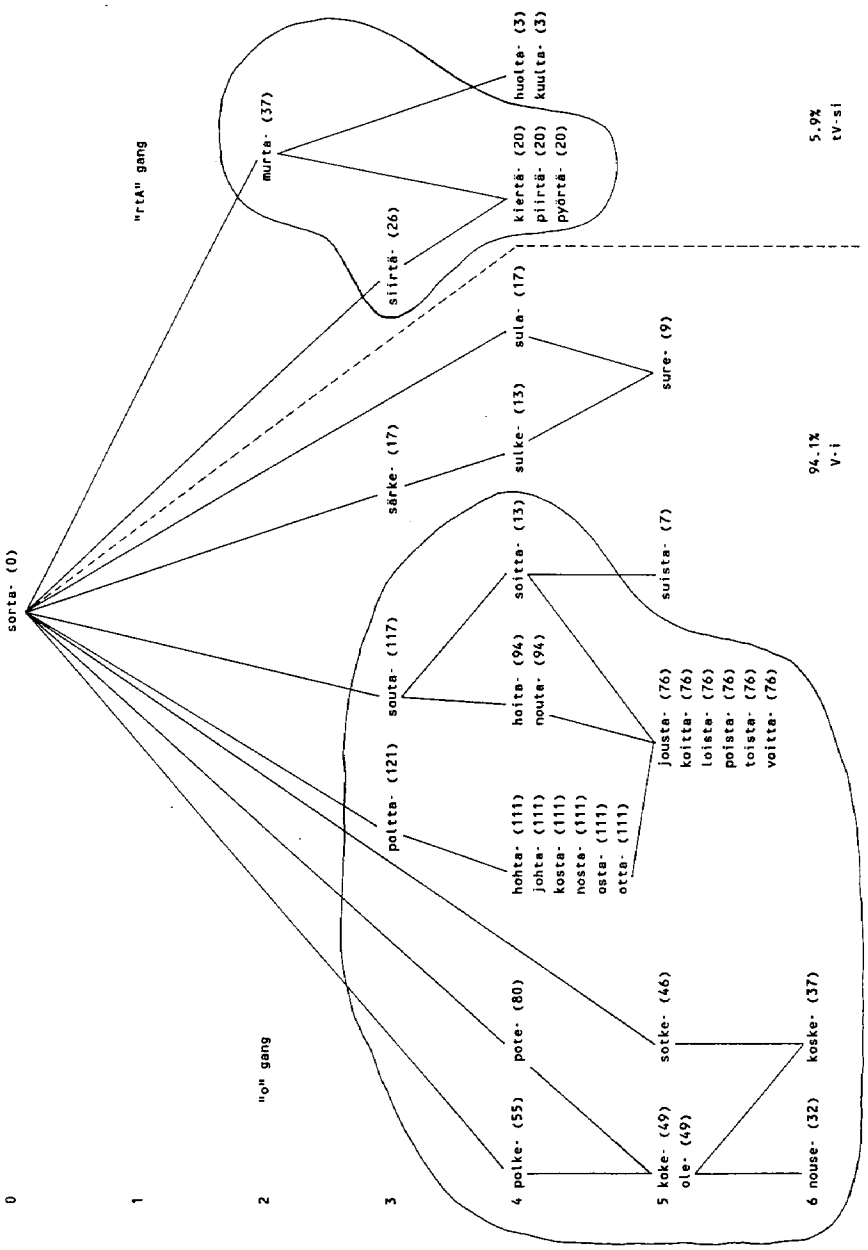

Figure 5.

13. A Procedural Alternative: Neural Networks

The analogical approach is a *procedural* one. By this, I mean that it can only react (or make a prediction) in terms of a given context. There is no global description of the data. On the other hand, rule approaches are *declarative*. Past behavior is specifically described, and in practice the resulting description is used (usually in imprecise or non-explicit ways) to predict subsequent behavior.

An important procedural alternative to analogical modeling is neural networks (also known as activation-deactivation modeling, connectionism, and parallel distributed processing). Since both neural networks and analogical models are procedural in nature, they share a number of properties (such as robustness, gang effects, fuzzy boundaries, apparent rule governedness, no competence-performance distinction, and so on). A good deal of the appeal of neural networks is due to their procedural properties, but analogical modeling also shares these same procedural properties. Perhaps more enticing has been the neurological metaphor itself, despite the fact that all neural networks are implausible if not impossible as actual neurological models. In fact, the term "neural network" is a misnomer; quite correctly, Freeman and Skapura (1991:3) call these models "artificial neural networks". Researchers should not be deceived into believing that their neural networks have much neurological basis. But the real question is whether neural networks make the appropriate general predictions about the nature of learned behavior. There are a number of problems with neural networks that should make us seriously reconsider them as models for explaining language. Here I will discuss two of them (for other objections, see section 4.2 of Skousen 1989).

13.1. Probability Matching

There is a good deal of evidence that speakers can learn probabilistic rules – or at least behave as if they have learned such rules. William Labov and his colleagues have provided many examples of language variation that cannot be reduced to deterministic behavior. Moreover, there are many psychological experiments that have demonstrated the ability of adults, children, and even laboratory animals to produce probabilistically – and with the same output probability as the input probability (Skousen 1989:82-85). Analogical modeling directly predicts that if the given context actually occurs, then the output probability will equal the input probability (with some variation for imperfect memory). On the other hand, neural networks do not readily allow for probabilistic behavior. A trained neural network typically activates the same outcome

every time for the same given context. It is difficult to get a specific outcome to be activated probabilistically, and if so it is virtually impossible to get it to be activated at a relative frequency consistent with an input probability. This problem of probability matching is a serious one and cannot be ignored.

13.2. Training the network

A number of problems have arisen in constructing and using neural networks. One problem is that the network can get stuck, so to speak, in a local minimum – and will not even know that it is stuck. The designer of the network must therefore devise various strategies so that local minima may be avoided.

Another disturbing problem has been the inordinately long time and numerous computer calculations necessary to train a network, even to learn simple behaviors. Consider the notorious case of XOR (the exclusive *or*), which has two input variables (each having two possibilities, 0 and 1) and a simple outcome, *T* or *F*. If both input variables agree, then the outcome is *F*; if both input variables disagree, then the outcome is *T*:

1	1	*F*
1	0	*T*
0	1	*T*
0	0	*F*

Training a neural network to learn this simple logical behavior takes some work. With only one hidden unit, thousand of data representations are necessary to train the network. Even with multiple hidden units, hundreds of data representations are necessary (Rumelhart and McClelland 1986: 330-334). And, of course, if there are no hidden units, it is impossible to learn the XOR behavior (Freeman and Skapura 1991:24-27).

The analogical approach to learning the exclusive *or* is very simple. The model simply has to remember at least one input example for each of the four possible cases (11*F*, 10*T*, 01*T*, and 00*F*). Given 11, the analogical model will immediately predict the outcome *F* from the example 11*F*. The data itself contains the information. No training is necessary. There are no local minima, hidden units, or extensive calculations.

Another important learning difference is that the analogical model makes each prediction from scratch. In this way the data set can be manipulated. We can readily study the effects of systematic change on

predicted behavior. Unfortunately, neural networks, once trained, cannot adjust to learn new kinds of input, but will collapse into predicting nonsense. In order to deal with input changes, the whole network has to trained all over again. This "catastrophe problem" is a serious challenge for neural networks as models of language. Linguistic change is continually occurring, and speakers adjust to it without having to learn the whole language over again.

14. Problems with Procedural Approaches

Procedural approaches do not seek to partition the contextual space or explicitly determine which variables are significant. As a consequence, there are a number of problems that confront both analogical models and neural networks. Here I will briefly discuss a couple of these problems.

14.1. Variable Specification

In developing procedural models, one common question is: What variables shall we specify (in either the data set or in the neural network)? Not only must we determine how to represent the variables, but there is also the question of just how many variables can we specify. There is undoubtedly some limit, perhaps based on short term memory. In the analogical modeling that I have done, computational limitations have restricted my data sets to twelve variables, sufficiently large enough to get interesting results for some problems, but obviously inadequate for a complete theory of language.

The number of variables has important consequences for computer processing. With sequential processing, each added variable basically doubles the processing time as well as the memory (hardware) requirements. This is the familiar algorithmic problem of exponential explosion. With parallel processing, the processing time can be reduced to a linear function of the number of variables, which is much more reasonable. Still, the hardware requirements remain considerable, although the increase is not quite exponential (Skousen 1989:137-139).

A second aspect of variable specification deals with the type of variables. Thus far nearly all examples of analogical modeling as well as neural networks have dealt with problems in phonetics, orthography, phonology, and morphology. In these cases we have a fairly good idea of what the variables should be like. Yet even here problems abound. And given our current knowledge, our ability to specify more abstract variables – sociological, syntactic, and semantic – remains problematic.

One important problem of variable specification deals with the question of representing sequential variables. Solving this problem will be particularly important for developing procedural models for syntactic relationship. Not only will we need to deal with questions of simple ordering, but also relationship that involve waiting (such as pronominalization, binding, and closure).

14.2. Language in Time

The syntactic question leads us to a second unresolved problem. Thus far most work in procedural approaches has dealt with predicting or interpreting one outcome at a time. Yet language obviously occurs in time and involves the prediction or interpretation of overlapping sequences of sounds, words, phrases, sentences, and discourse elements. In other words, how do we deal with language as it is produced in time – as speech in process? In this paper I have discussed examples of predicting the indefinite article in English and the past tense in Finnish, but obviously speakers do not generally predict such forms in isolation. Language processing does not devote itself to solving one outcome at a time – and probably not even as a well-defined sequence of outcomes.

15. Current Work in Analogical Modeling

The basic introduction to the analogical approach is *Analogical Modeling of Language* (Skousen 1989). The mathematical basis for analogical modeling (as well as rule-based systems of description) is found in a more extensive work entitled *Analogy and Structure* (Skousen 1992). I am currently doing research for a book (tentatively entitled *Natural Statistics*) that will describe the mathematical equivalence of natural statistics and traditional statistics.

A number of other research programs are currently under investigation. Bruce Derwing and I are applying the analogical model to the English past tense and testing it against a large database of experimental results collected by Derwing over a number of years. In addition, Derwing plans to do some experimental testing of my predictions about the past tense in Finnish.

Steve Chandler, of the University of Idaho, is also investigating the possibility of doing analogical modeling in terms of neural networks. Chandler has asked the important question of whether the analogical approach can be considered a type of neural network. Under a broadly conceived system of connections, some aspects of deriving the analogical set may be possible from an activation-deactivation perspective. But in

order to do this, a number of strict limitations would be necessary. Instead of positing a single node to represent a specific outcome, individual nodes would represent specific exemplars. The connections between nodes would probably have to be a binary one (either off or on), and the resulting analogical set would need to permit a number of exemplars to remain activated (thus achieving probability matching).

Chandler has brought to my attention an interesting paper (Burton 1990) that discusses the possibility of an exemplar-based approach to learning by dividing up cerebral processing into "discrete temporal gates". Burton refers to these exemplars as "episodes" and argues that subconscious learning proceeds from these episodes. The neurological possibilities for an exemplar-based procedural approach is an intriguing idea and worthy of continued research.

Address of Author

Royal Skousen: Department of English, Brigham Young University, Provo, Utah, 84602 USA, e-mail: royal_skousen@byu.edu

Footnotes

[1] An earlier version of this paper was delivered at the 21st Annual Linguistics Symposium, "The Reality of Linguistic Rules", The University of Wisconsin at Milwaukee, 12 April 1992.

References

BURTON P.G. (1990), "A search for explanation of the brain and learning: Elements of the psychonomic interface between psychology and neurophysiology", *Psychobiology* 18:119-194.

FREEMAN J.A. & SKAPURA D.M. (1991), *Neural Networks: Algorithms, Applications, and Programming Techniques*, Reading, Massachussetts, Addison-Wesley.

RUMELHART D.E. & MCCLELLAND J.L. (1986), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 1, Cambridge, Massachussetts, MIT Press.

SADENIEMI M. ed. (1973), *Nykysuomen sanakirja*, Porvoo, Finland, Werner Söderström.

SKOUSEN R. (1989), *Analogical Modeling of Language*, Dordrecht, The Netherlands, Kluwer.

SKOUSEN R. (1992), *Analogy and Structure*, Dordrecht, The Netherlands, Kluwer.