Expanding Analogical Modeling into a General Theory of Language Prediction
Royal Skousen
Max Planck Institute, Leipzig
23 September 2006

1. The Core Theory

In Analogical Modeling (AM), I distinguish between the core theory and its application to language. I begin here by briefly defining the basic system of AM. The goal is to predict the **outcome** for a set of conditions referred to as the **given context** (sometimes the given context is referred to as the **test item**.) From the given context, we construct more general versions of that context, which we refer to as **supracontexts**. Our goal is to predict the behavior (or outcome) of the given context in terms of the behavior of its supracontexts. The source for determining those behaviors comes from a **dataset of exemplars**; for each exemplar in the dataset, the outcome is specified. These exemplars, with their own specifications and associated outcomes of behavior, are assigned to the various supracontexts defined by the given context. Supracontexts that behave uniformly (referred to as **homogeneous** supracontexts) are accepted, with the result that exemplars contained within the homogeneous supracontexts can be analogically used to predict the behavior of the given context. The exemplars found in nonuniformly behaving supracontexts (referred to as **heterogeneous** supracontexts) cannot be used to make the analogical prediction for the given context. The term *nonuniformity* means that a heterogeneous supracontext has a plurality of subcontexts and a plurality of outcomes (that is, exemplars within the supracontext not only have different outcomes but they are also found in different subspaces of the contextual space). Finally, the relative probability of using a homogeneous supracontext is equal to the square of its relative frequency, while the probability of using a heterogeneous supracontext is zero. (For a basic introduction to AM and how it works, see Skousen, Lonsdale, and Parkinson 2002:12-22 or Skousen 2003.)

One simplified way to look at AM is in terms of traditional **rules**, where the term *rule* basically stands for the supracontext and its associated behavior. In trying to predict the behavior of the given context, we consider all the possible rules that could apply. We eliminate those rules that behave nonuniformly (that is, the heterogeneous rules). All uniformly behaving rules (the homogeneous rules) are then applied, with the probability of applying a given homogeneous rule equal to the square of its relative frequency. One important aspect of AM is that each rule's homogeneity can be determined independently of every other rule. This property of independent determination of uniformity means that we can examine a rule's uniformity without having to determine whether any subrule (that is, any more specific version of the rule) behaves differently.

AM is computationally intensive. For each variable added to the specification of a given context, both the memory requirements and the running time doubles (so if there are n variables in the given context, the memory and time are of the order 2^n). This problem of exponential explosion has been theoretically solved by redefining AM in terms of Quantum Analogical

Modeling (QAM), a quantum mechanical approach to doing AM. The main difference is that everything is done simultaneously in QAM, in distinction to the sequential application that AM is forced to follow. Still, the same basic procedure is followed, only the system of rules (or supracontexts) is now treated as a quantum mechanical one:

(1) all possible rules for a given context exist in a superposition;
(2) the exemplars are individually but simultaneously assigned to every applicable rule;
(3) the system evolves so that
    (a) the amplitude of every heterogeneous rule equals zero, while
    (b) the amplitude of each homogeneous rule equals its relative frequency of occurrence;
(4) measurement or observation reduces the superposition to a single rule
    where the probability of it being selected is equal to its amplitude squared
    (that is, equal to its relative frequency squared).

See See Skousen 2002:319-346 for an introductory essay on treating AM as a quantum mechanical system. For a complete discussion of how QAM works, see Skousen 2005.

One notices here that nothing in the core of AM specifies how AM is to be applied to language. All such language applications have their own linguistic assumptions, and it is an open question not directly related to AM itself on what those assumptions should be. But by choosing various assumptions and seeing what kinds of predictions AM makes about the behavior, then by comparing the predicted behavior to the actual behavior, we can assess the empirical validity of those linguistic assumptions.

A similar situation exists in quantum mechanics, which seems appropriate to bring up here since QAM itself is a quantum mechanical system. As explained by Charles Bennett, there is a "set of laws" (like the Ten Commandments, as he puts it) that form the basics of quantum mechanics (QM), but QM has to be applied in order to serve as a theory of physics: "For most of the 20th century, physicists and chemists have used quantum mechanics to build an edifice of quantitative explanation and prediction covering almost all features of our everyday world." The core theory is actually very simple, but the resulting edifice is complex and evolves. Yet in all instances, QM involves applying the core theory and making hypotheses regarding the underlying physical system, which if the resulting application of the theory works, we accept the hypotheses as representing, in some sense, physical reality. For a pictorial representation of this point, see Charles H. Bennett, "Quantum Information Theory" (in Hey 1999:177-180).

AM is a general theory of predicting classification and is not restricted to linguistic problems per se. For instance, one can use AM to predict various kinds of nonlinguistic outcomes, such as determining whether different mushrooms are poisonous or not, providing medical diagnoses based on symptoms and lab tests, and predicting party affiliation on the basis of voting patterns. For various examples, see Deryle Lonsdale, "Data Files for Analogical Modeling", in Skousen, Lonsdale, and Parkinson, 2002:349-363.

2. Basic Structural Types

The current AM computer program treats the *n* variables defined by a given context as *n* independent variables, which means that any linguistic dependencies between the variables must be built into the variable specifications. Very seldom can we construct cases where there are no linguistic dependencies in the variable specification (although there will always be behavioral dependencies between the variables). One possible example of linguistic independence of variables involves the social features for specifying terms of address in Arabic (see Skousen 1989:97-100), which has social variables like age of speaker, gender of speaker, and social class relationship. Very often the linguistic task involves strings (in phonology) or trees (in morphology and syntax). The general approach in Skousen 1989 was to treat strings of characters as variables for which the position was specified. For instance, in predicting the spelling of the initial /h/ sound in English words (as either *h*, *wh*, or *j*), positional aspects were included in the definition of each variable, such as "the **first** vowel phoneme" and "the phoneme that **immediately precedes** the **third** vowel". This kind of variable specification allows the AM computer program to make the analysis, but it is not realistic since it requires that everything be lined up in advance so that the strings can be compared. Cases of metathesis or identical syllables in different positions are ignored, nor can they be readily handled in such a restricted version of AM.

These kind of specifications have led to the use of zeros for variables. For instance, if a word has only one syllable, then the nonexistent second and third syllables are marked with zeros. But then there is the question of how to specify such nonexistent syllables. If we mark the nuclear vowel for such a syllable with a zero, do we also mark the syllable's onset and coda with zeros, even though those zeros are redundant? One possibility is to refer to such predictable zeros as redundant variables and to ignore them when making analogical predictions (the general way of proceeding in Skousen 1989). If all zeros (both essential and redundant) are counted, then there is the possibility that the analogical prediction will be overwhelmed by excessively specified zeros. But there is also the possibility that we may want to count all the zeros (for some discussion of this issue, see Skousen, Lonsdale, and Parkinson 2002:40-42). The important point here is that the problem of the zeros results from trying to account for strings as if they were composed of unordered symbols. One major issue that linguistic applications of AM must deal with then is how to treat strings and trees as they actually are rather than trying to define them as sets of independent variables. In the remainder of this section, I outline several different approaches to structures that one would want to use in linguistic analyses. It should be pointed out, however, that the current AM computer program has not yet been revised to handle these structures directly.

2a. Strings of characters

A more reasonable approach for a string of characters would be to allow any possible sequence of substrings of a given string to count as a supracontext. For instance, if the given string is *abc*, the supracontexts would include examples like **abc, a*b*c, *ab*c*, a*b,* and **c,* where the asterisk stands for any string, including the null string. Thus the supracontext **abc*

would include any string ending in *abc* while **ab*c** would include any string containing *ab* followed by *c*. The most general supracontext would be simply * (that is, this supracontext would contain all possible strings, including the null string). For *n* characters in a given string, there will be a total of $2 \cdot 3^n$ supracontexts for which we will need to determine the heterogeneity, but (as already noted) this can be done independently for each supracontext by determining its heterogeneity with respect to the outcomes and the subcontexts for the data items assigned to that supracontext. The total number of supracontexts for this kind of string analysis is also exponential (like the 2^n for when the *n* characters are all independent variables), but that number ($2 \cdot 3^n$) increases at a greater exponential rate.

2b. Scalar variables

The AM computer program assumes that the variables specified by the given context are categorical and discrete. The question then arises of how to deal with scalar variables, ones that represent degrees of a property. Scalars can be mathematically treated as real numbers, but this leads to extraordinary problems with the number of possible supracontexts since theoretically every possible real number interval could count as a supracontext, which ends up defining a nondenumerably infinite set of supracontexts. I would propose, instead, that continuous scalars be analyzed as a sequence of finite intervals (that is, we will quantize the scalar). Having made that decision, we can then decide how to determine the supracontexts for a given sequence of finite intervals.

As an example, consider how we might apply this quantization to the problem of voicing onset time and the ability of speakers to predict whether a given stop is voiced or voiceless. Our task is to model how speakers interpret artificial stops with varying lengths of nonvoicing after the release of the stop. In this case, the data comes from experiments testing the ability to distinguish between /b/ and /p/ in English. The variables center around the problem of dealing with a time continuum. In applying AM to this problem, I assume that time should not be treated as a real number line. Instead, time will be broken up into a sequence of finite intervals of time, all equal in length, as described in Skousen 1989:71-76. Given an overall length of about 50 msec between instances of /b/ and /p/, let us break up this overall length into 5 intervals of 10 msec each, so that instances of voiced stops are represented as *xxxxx* and voiceless stops as *ooooo*. For simplicity of calculation, I will assume that there is in the dataset but one occurrence of each voiced stop, /b/ and /p/. The question then becomes: What are the supracontexts for the intermediate but nonoccurring given contexts (namely, *oxxxx, ooxxx, oooxx, oooox*)? I will here consider three possibilities for the supracontexts defined by the particular given context *ooxxx*:

(1) We treat each single continuous sequence of intervals as a possible supracontext. The number of homogeneous supracontexts, in this case, will be quadratic – namely, $n(n+1)/2$:

given context	/b/ *outcome*	/p/ *outcome*	*probability*
xxxxx	15	0	1.000
oxxxx	10	1	0.909
ooxxx	6	3	0.667
oooxx	3	6	0.333
oooox	1	10	0.091
ooooo	0	15	0.000

(2) We treat each *o* and *x* and its position as an independent variable (this is how the problem is treated in Skousen 1989:71-76). This means that any subset of the 5 variables will define the possible supracontexts. The number of homogeneous supracontexts will be exponential (to the scale of $2^n - 1$), which means that in comparison with the previous case, the shift in predictability will be sharper. We get the following predicted chances for /p/ and /b/:

given context	/b/ *outcome*	/p/ *outcome*	*probability*
xxxxx	31	0	1.000
oxxxx	15	1	0.938
ooxxx	7	3	0.700
oooxx	3	7	0.300
oooox	1	15	0.063
ooooo	0	31	0.000

(3) Finally, we treat the sequence of intervals as a string and permit any set of nonoverlapping substrings to serve as a distinct supracontext. In this case, the shift in predictability will be sharper than in the second case (but also exponential) since the number of homogeneous supracontexts will have the exponential factor $2(3^n - 1)$ rather than the $2^n - 1$ of the second case:

given context	/b/ *outcome*	/p/ *outcome*	*probability*
xxxxx	484	0	1.000
oxxxx	160	4	0.976
ooxxx	52	16	0.765
oooxx	16	52	0.333
oooox	4	160	0.024
ooooo	0	484	0.000

For each of these three cases, we can determine which interval length allows for the best fit for the actual experimental results for predicting /b/ versus /p/ (see Lisker and Abramson 1970, cited in Skousen 1989). For an overall interval of 50 msec, we get the following:

type of analysis	*number of homogenous supracontexts*	*number of intervals*	*length of interval*
(1) a single continuous substring	$n(n+1)/2$	10	5 msec
(2) *n* independent variables	$2^n - 1$	7	7 msec
(3) any nonoverlapping sequence of substrings	$2(3^n - 1)$	5	10 msec

Lehiste 1970 (cited in Skousen 1989) provides evidence that speakers can distinguish between sound durations differing as little as 10 milliseconds, which means that the last case (which defines the supracontexts as any nonoverlapping sequence of finite intervals) is the one that best corresponds with experimental results for humans trying to distinguish between artificial versions of /b/ and /p/ in terms of voicing onset time.

2c. Unordered hierarchical structures (branching hierarchical sets)

This kind of structure is found in semantics. The supracontexts for a given hierarchical set are subsets that generalize by moving up the hierarchy, thus accounting for hyponymy. Semantic variables (or features) defined for lower, more specific subsets may be ignored in higher, more general subsets.

The need for localized restrictions on the use of semantic features is well exemplified in an attempt to analyze and predict the behavior of Chinese classifiers, found in some unpublished work by my colleague Dana Bourgerie, presented at the 2000 Analogical Modeling conference at Brigham Young University ("An Analysis of Chinese Classifiers: Issues in Dealing with Semantic Variables in the AML Framework"). Some of the classifiers examined by Bourgerie were:

ge	general classifier for people, round things, and things of indeterminate form
zhang	for open flat things (maps, tables, tickets, etc.)
tiao	for long, thin things (fish, leg, boat, cucumber, a long bench, etc.)
zhi	for long, branch-like things (e.g. pen, gun, candle, etc.)
ba	things with handles (e.g. umbrellas, swords, etc.)
jian	mostly for rooms
ben	for bound things such as books

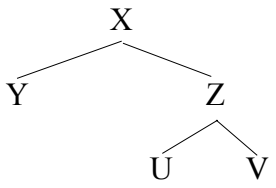
The need for an analogical model results from a great deal of variability in actual usage between speakers in selecting the appropriate classifier as well as the extension of classifiers to new objects.

Bourgerie's variables were as follows: relative size (*s, m, l*), flat (+, –), long (+, –), narrow (+, –), three-dimensional (+, –), handle (+, –); every instance of usage involving a classifier in his dataset was defined in terms of these specific variables. Given what we know now, the size variable should have been converted to a discrete scalar (something like – –, –+, and + + to stand for small, medium, and long, respectively); I will make that conversion here to simplify the description. In other words, the semantic description of every noun in the dataset can be analyzed as a sequence of pluses and minuses. In Bourgerie's preliminary work, pluses and minuses were assigned in all cases. For example, *ba* is expected for objects with a handle, even though other items with handles, such as a gun, take *zhi*. On the other hand, some objects, such as boat or long bench (which take the classifier *tiao*), do not ordinarily not have handles, yet '–handle' was assigned to this classifier. And finally for some nouns referring to people (which take the classifier *ge*), handles would seem implausible, although one could imagine it! Implausible or not, '–handle' was assigned to words taking the most general classifier. And similar overloading of minus-valued variables occurred for other specific classifiers. The overall result was that the classifier *ge*, being the most general classifier, had more minuses for the words assigned to it – and especially more minuses than words assigned to the other (more specific) classifiers.

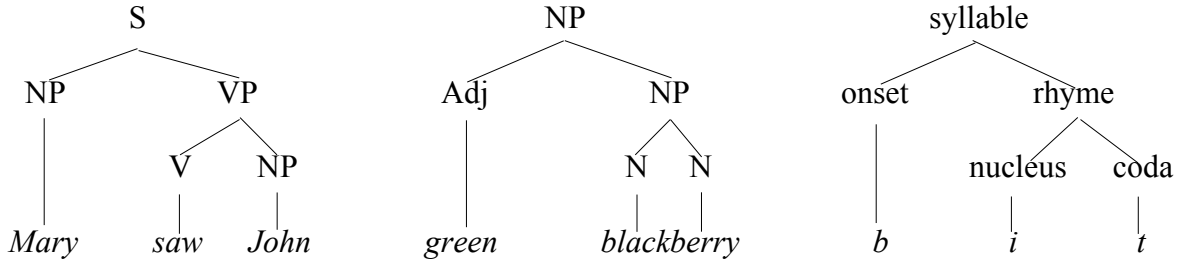
The problem with assigning '–handle' (and similarly for other specific variables) to all the nouns in the dataset is that when predicting the classifier for any given word, the minuses dominate, with the result that the general classifier *ge* consistently swamps the prediction, even when we are predicating an item close to words that take one of the more specific classifiers. To get the right results, we need to restrict '+/– handle' to smaller groups of words where they characteristically are found. So we may mark kitchen utensils as to whether or not they have a handle, but not the refrigerator, oven, sink, dishwasher, counters, tables, chairs (although they could have them). Where the handle helps to distinguish between closely associated objects that are named differently, the variable should be assigned, but otherwise not. A table could have a handle, but such a table doesn't have a different name, so we do not specify it as a variable in such cases. The vast majority of words could always be marked as '–handle' (such as a cloud, a tree, a lake, a newspaper, philosophy, war, etc.), but AM shows that we cannot semantically analyze every word as plus or minus for every possible semantic variable. This may seem obvious: Would we really want to mark virtually every object in the world as '–human'? Semantic variables are defined within only certain restricted domains. In applying AM to Chinese classifiers, Bourgerie marked every word in the dataset as either '+handle' or '–handle' and soon discovered that such a decision clearly made the wrong predictions.

2d. Ordered hierarchical structures (*trees*)

Ordered hierarchical structures obviously have both order and hierarchy and are commonly referred to as trees. Given a particular tree as a given context, the supracontexts are defined as subtrees of the given tree. For instance, our given contexts and the data items may best be represented as trees and we may wish to predict some behavior given such a tree. The following simple right-branching structure is of interest in many different situations:



We find uses of it in specifying syntax, morphology, and syllable structure:



In attempting to predict some outcome based on the pronunciation for the last item, *beet* /bit/, we could restrict the supracontexts for the given context (namely, the tree itself) to combinations of categories that occur only at the same level in the tree; for instance, we could examine all syllables with the same onset, or with the same rhyme (nucleus and coda), or with the same nucleus, or with the same coda – but not with the same onset and the same nucleus or with the same onset and the same coda since those categorical combinations do not occur at the same level in the tree. Of course, we would only want to do this if there was evidence that such a restriction on supracontextual construction would predict language behavior. In other words, the decision is more an empirical one than one that impinges on the question of whether AM is a correct theory.

We get similar hierarchical problems in specifying distinctive features. As discussed in Skousen 1989:53-54, we cannot treat distinctive features as if they are independent variables. Suppose we compare *beet* /bit/ with two possible words, each of which differs from /bit/ in three distinctive features. If we treat this problem as a set of 12 variables, the distance between *beet* and *bought* is the same as between *beet* and *mid*:

(a) *three-feature difference restricted to one phoneme:*

	consonant	vowel	consonant
/bit/	oral stop labial voiced	**spread high front** tense	oral stop alveolar voiceless
/bɔt/	oral stop labial voiced	**round low back** tense	oral stop alveolar voiceless

(b) *three-feature difference spread across three phonemes:*

	consonant	vowel	consonant
/bit/	**oral** stop labial voiced	spread high front **tense**	oral stop alveolar **voiceless**
/mɪd/	**nasal** stop labial voiced	spread high front **lax**	oral stop alveolar **voiced**

Yet experimental evidence from perceptual studies show that speakers perceive *beet* and *bought* as phonetically close, while *beet* and *mid* are not especially close (see Derwing and Nearey 1986, cited in Skousen 1989). If we treat distinctive features as independent variables, we incorrectly predict an equality of phonetic similarity for this example. One way to correct this would be to define the given contexts in terms of phonemes and basic syllable structure, which would mean that there is only one difference between *beet* and *bought*, but three between *beet* and *mid* (this is how it is done in Skousen 1989). But another possibility would be to define distinctive features for only phonemic nodes within syllable tree structures, thus restricting feature similarity to apply only at isolated places in the tree.

3. Control Over the Analogical Set

The general theory of analogical modeling (AM) allows for various ways of using the analogical set to predict outcomes (although the quantum version of it, QAM, does not). Here I review this aspect of AM.

3a. Reacting to a previous prediction

One important point is to recognize that analogical modeling allows for the ability to reexamine a given analogical set or to re-determine it under various conditions. A speaker may, for instance, produce a particular outcome, but then not like the results and so produce a different outcome. The speaker does not get caught in an infinite loop, continually producing, say, the most favored outcome or randomly producing outcomes, thus leading to the repetition of the more frequent outcomes. Consider, for instance, the following two examples from my own children's speech (cited in Skousen 1989:85-86):

Nathaniel (5 years, 10 months):

> Looking at a picture of the Grand Canyon, Nathaniel keeps trying to produce the plural *cliffs*: /klɪˈftəz/, /klɪfs/, /klɪvz/, /klɪfs/
>
> > Note that Nathaniel's sequence of productions is not constantly repetitive (as if it were /klɪfs/, /klɪfs/, /klɪfs/, /klɪfs/, ...).

Angela (6 years, 10 months)

> The possessive form *Beth's* is pronounced first as /bɛs/ and then immediately followed by /bɛˈθəz/.
>
> > Angela: How do you add the *s* to *Beth*? It's hard to say. How do you say it?
> > Royal: I say /bɛs/ [bɛs:].
> > Angela: I say /bɛθ/ like *Beth house* /bɛθ haus/.

Note that Angela produced a sequence of three different possibilities: /bɛs/, /bɛˈθəz/, /bɛθ/.

Angela (7 years, 11 months)

The plural form *ghosts* is pronounced initially as /gousts/, then as /gous/, and is finally followed by the question "How do you pronounce that?"

Similarly, suppose we have a nonce word (written out) and ask someone to pronounce it; then no matter what they say, we say that it's wrong and ask for an alternative pronunciation. Our subjects do not go into an infinite loop, instead they will typically produce a sequence of different responses. As example is the nonce word YEAD, which might be pronounced alternatively as /yid/, /yɛd/, /yeid/.

For each new prediction, we could let the analogical set be re-determined from scratch but with all data items having the forbidden outcome eliminated so that those exemplars will not play a role in constructing the analogical set, especially since the original analogical set may only provide one possible outcome. Or maybe one has a choice: Try the original analogical set first; if that fails, then revert to re-determining it by omitting the forbidden outcome.

3b. Random selection versus selection by plurality

Another aspect dealing with control over the analogical set is the choice between random selection of an outcome and selection by plurality (discussed in Skousen 1989:82-85). Psycholinguistic experiments show that speakers of all ages can reproduce probabilistic behavior by applying random selection to the analogical set. But as speakers grow older, by about age 8, they are also able to select the most frequent outcome, especially when they expect or want to make some gain from the choice of outcome. It can also be shown that if the choice involves some loss, then the most advantageous decision is to choose the least frequent outcome (discussed in Skousen 1992:357-358). The ability to select by plurality would apparently require some kind of sampling or analysis of the analogical set, perhaps as it is being determined.

3c. Restricting morphological extension

Another issue involving restrictions on the use of the analogical set asks whether there are any limits besides heterogeneity in preventing the overuse of analogy. Consider, for instance, the analogical prediction of the past tense in English for the verb *be*. The question here is whether the verb *see* (with its exceptional past-tense form *saw*) can be used as an exemplar in predicting the past-tense form for *be*:

/si/ : /sɔ/ :: /bi/ : /bɔ/ (that is, *see* : *saw* :: *be* : *baw*)

This analogical extension seems highly unlikely. One might argue that such an analogy is difficult simply because the chances of forgetting the past-tense *was*/*were* for the very frequent verb *be* are virtually negligible. But the question still remains: Is *baw* even possible? And if so, is there any way besides appealing to heterogeneity to restrict the applicability of *saw*? Here heterogeneity may not work since *see* is such a close neighbor to *be*, at least close enough to allow it to analogically apply to *be*.

Since we know the analogical set can be examined prior to using it, perhaps the speaker can reject an unrecognizable past-tense form. One could argue that the analogical set provides only results, not how those exemplars are derived. The analogical *baw* could therefore be possible, but at the same time unrecognizable, thus one could simply avoid using it. A similar case involves verbs of the form CX-Cɔt:

alternation	*example*	*extension*
ing-ɔt	bring-brought	sting-stought
ink-ɔt	think-thought	drink-drought
ach-ɔt	catch-caught	latch-lought
ai-ɔt	buy-bought	try-trought
ich-ɔt	teach-taught	reach-rought
ik-ɔt	seek-sought	tweak-twought

Is heterogeneity sufficient to prevent any of the these analogies from applying? Probably not. But these analogies could nonetheless be rejected by speakers since the resulting past-tense forms are unrecoverable – that is, speakers are unable to determine what verb the past-tense form stands for. A past-tense prediction like *stought* would imply only that the analogical present-tense verb form began with *st*.

One could propose that unique alternations can never be extended analogically, but this is definitely false. We have, for instance, analogical extensions based on the noun *ox* and its uniquely exceptional plural form *oxen* (thus *axen* for the plural of *ax* and *uxen* for the nonce *ux*). But note that in these cases the singular forms *ax* and *ux* are recoverable from *axen* and *uxen*. The question may not be one of uniqueness, but rather recoverability.

4. Specifying the Variables

One important aspect of AM is that we not restrict our analysis to just the important or crucial variables. We need to include "unimportant" variables in order to make our predictions robust. Consider, for example, the indefinite article *a*/*an* in English. Knowing that the following segment, whether consonant or vowel, "determines" the article (*a* for consonants, *an* for vowels), one could specify only the syllabicity of the following segment and thus predict *a*/*an* without error. Basically, we would be specifying a single rule analysis for the indefinite article. Yet in modeling the behavior of the indefinite article, AM specifies in addition the phonemic

representation for that first segment in the following word as well as the phonemes and syllabicity for other segments in that word, supposedly unimportant variables. But by adding these other variables, AM is able to predict several behavioral properties of the indefinite article: (1) the one-way error tendency of adult speakers to replace *an* with *a* (but not *a* with *an*); (2) children's errors favoring the extension of *a*, but not *an*, such as "a upper", "a alligator", "a end", "a engine", "a egg", and "a other one"; (3) dialects for which *an* has been replaced by *a*, but not the other way around. In other words, the "unimportant" variables are crucial for predicting the fuzziness of actual language usage (for some discussion of these properties, see Skousen 2003). Finally, another important property is that AM can predict the indefinite article even when the first segment is obscured (that is, when one cannot tell whether that segment is a consonant or a vowel). In such cases, the other variables are used to guess the syllabicity of the obscured segment, thus allowing for the prediction. In other words, AM allows for robustness of prediction. If we assume a symbolic rule system with only one rule (one based on the syllabicity of the first segment), then no prediction is possible when that segment is obscured. For additional discussion of the robustness of AM with respect to the indefinite article, see Skousen 1989:58-59.

Specifying "unimportant" variables also allows for cases where the preferred analogy is not a nearest neighbor to the given context, but is found in a gang of homogeneous behavior at some distance from the given context. An important example of this occurs in predicting the past tense for the Finnish verb *sortaa* 'to oppress'. Standard rule analyses of Finnish as well as nearest neighbor approaches to language prediction argue that the past tense for this verb should be *sorsi*, whereas in fact it is *sorti*. Yet when AM is applied to predicting the past tense in Finnish, it is able to predict the correct *sorti*, mainly because AM indirectly discovers that the *o* vowel is the "crucial" variable in predicting the past tense for this verb. In previous analyses (typically based on the historically determined "crucial" variables), the *o* vowel was ignored. But in AM, by specifying variables (both "important" and "unimportant") across the whole word, was able to make the correct prediction for this "exceptionally behaving" verb. For a complete discussion of how AM solves the problem of *sortaa*, see Skousen, Lonsdale, and Parkinson 2002:27-36.

4a. Varying the granularity of prediction

Computationally, there is a need to limit the number of variables. The current AM program can handle up to 60 variables, although the processing times can become quite long whenever there are more than 40 variables. The problem here is that the actual computer program is sequential and does not simultaneously run an exponential number of cases (as the proposed quantum computer would). Even the parallel processing provided by standard supercomputers does not appear to be capable of eliminating the fundamental exponential explosion inherent in AM. Presumably there are also empirical limitations on the number of variables that are processed. In other words, there will be a degree and type of granularity that results from how many and which variables are selected. Ultimately, we have to select the variables, but we want to judiciously select variables in a principled way that will, at the same time, allow for general applicability. In Skousen 1989:51-54, I suggest that enough variables be selected so that each exemplar in the dataset is distinguishable or recognizable. It is this property that argues for

specifying more than the first segment of the following word in predicting the indefinite article *a*/*an*. Or in the case of *sortaa*, we specifying variables across the entire word (thus including the *o* vowel). Another suggestion is that proximity to the outcome should be accounted for. For instance, in trying to predict the ending for a word, if we want to provide variables for the antepenultimate syllable, we should also provide variables for the penultimate and ultimate syllables.

4b. Avoiding inappropriate variables

There are undoubtedly some variables that are inappropriate, either conceptually or empirically. For instance, in predicting the negative prefix for adjectives in English, we could consider specifying the etymological source of adjectives since there is some correlation (although imperfect) between selecting the Latin negative prefixes *in-, il-*, *ir-,* and *im-* for words of Latin origin and the invariant Germanic negative prefix *un-* for words of Germanic origin. It turns out that such an etymological variable will have some influence in helping to predict the correct prefix (but not as much as one might suppose since historically these prefixes have been extended to words of different etymological background). From a conceptual point of view, in modern English, we cannot claim that speakers know the etymologies of the adjectives (although this may have been true for some educated speakers earlier in English when the influx of Latin vocabulary was in its beginning). For further discussion of this issue, see Chapman and Skousen 2005:341-342.

As an example of an empirical restriction on variables, consider whether multi-syllable words should be specified in terms of stress pattern or number of syllables. For instance, in predicting the past tense for Finnish verbs, Skousen 1989:101-104 used a restricted dataset: two-syllable verbs ending in a non-high, unrounded vowel (*e*, *ä*, or *a*). The results were very accurate in predicting speakers' intuitions as well as historical and dialect development. But extending the dataset to the entire verb system was much more difficult until it was realized that the variables should be specified in terms of stress pattern rather than by number of syllables. This difference may seem surprising since stress is supposed to be fully predictable in Finnish (primary stress on the first syllable, secondary stress on alternating syllables according to syllable weight). Yet there is empirical evidence that Finnish speakers rely on stress rather than number of syllables. Consider the following two analyses of the Finnish illative ending (meaning 'into'), where the first analysis is based on counting the number of syllables, the second on the kind of stress placed on the last syllable:

number of syllables

one syllable, long vowel or diphthong	−hV$_i$n
two or more syllables	
long vowel	−seen
diphthong	−hV$_i$n
short vowel	−V$_i$n

stress

stressed, long vowel or diphthong	−hV$_i$n
unstressed	
long vowel	−seen
diphthong	−hV$_i$n
short vowel	−V$_i$n

(Here V$_i$ means that the stem-final vowel is copied.) There is basically no difference between these two analyses since primary stress is virtually always on the first syllable. The crucial distinction between the two analyses is brought out when we consider how Finnish speakers predict the illative for two-syllable loan words where the original primary stress on the final syllable has been maintained. And the answer is that they follow the stress-based analysis:

Rousseau	rusó:	rusó:hon
Bordeaux	bordó:	bordó:hon
Calais	kalé:	kalé:hen

But if these words were nativized, with stress on the first syllable, then speakers would produce illative forms like /kále:se:n/. This means that in specifying the variables for Finnish words, we need to provide information regarding the stress pattern, not the number of syllables.

4c. Weighting of variables

Now if we decide that we must specify the stress pattern, an important question arises: What is the strength of the stress in predicting the outcome? Is it the same as the individual phoneme? Consider, for instance, variables that might be specified for the syllable in Finnish (the 9 variables listed here are much like the ones used in Skousen 1989):

1 syllable-initial consonant (include 0 as a possibility)
  * a syllable-structure alternative:
    (1a) is there an initial consonant or not?
    (1b) if so, what is it?
2 the nuclear vowel: specify its phoneme
3 is there a second vowel or not?
4 if so, what is it?
5 is there a sonorant in the coda?
6 if so, what is it?
7 is there a obstruent in the coda?
8 if so, what is it?
9 what is the stress on the syllable? primary, secondary, none
  * a scalar alternative (10, 00, 01):
    (9a) is the stress primary?
    (9b) is there no stress?

If we follow the two alternatives (each marked with an asterisk), we have 11 variables, of which 4 deal with syllable structure, 5 specify the sounds (here the phonemes), and 2 the stress. If we analyze the phonemes into distinctive features, the number of variables specifying sounds would at least triple and probably overwhelm the analysis. Perhaps even as it is, the two variables dealing with stress may not be enough. Even worse would be specifying a single stress variable for the intonational contour of the entire word.

This problem becomes more acute when one specifies variables from completely different types of linguistic classification, say phonetic and semantic. Suppose we are trying to predict an outcome, say a grammatical gender, that is affected by the phonetics of the word as well as whether the word refers, say, to animates or non-animates. We set up say 10 or so variables for the phonetics of the last syllable (as a minimum). But then the question is: Do we assign just one variable to tell us whether the word refers to an animate or non-animate object? It is very doubtful that a single variable assigned to animacy will be strong enough to show the influence of that semantic class. Just doubling or tripling that semantic variable seems awfully arbitrary, although from a pragmatic point of view one could increase the strength of such a variable until one gets the right results! David Eddington did precisely that in his article “A Comparison of Two Analogical Models” when he considered the relative strength of phonemic variables versus morphological variables (Skousen, Lonsdale, and Parkinson 2002:148):

> Therefore, in addition to the phonemic information, morphological variables were included. For verbal forms, one variable indicated the person, and three identical variables indicated the tense form of the verb. Repeating a variable more than once is the only way to manipulate the weight of one variable or another prior to running the AM program. In essence, what this implies is that the tense form of the verb is considered three times more important that any single onset, nucleus or coda. In the AM simulation, the only significant difference that weighting this variable made was in the number of errors that occurred on preterit verbs with final stress. Fifty errors occurred without the weighting, in comparison to 27 when it was included three times.

And it should be remembered that this approach will not work if the variable being considered has no effect on the outcome. Dirk Elzinga (2006:766) reports that in using AM to predict the comparative for English adjectives, he used a morphological variable that specified whether the adjective was morphologically simple or complex, and he discovered that doubling, tripling, and quadrupling that morphological variable had no effect on the predictability of the outcome.

Obviously, we need a principled method of constructing variables so that the empirically-determined relative strength between classificatory types is naturally achieved.

5. Specifying the Outcomes

5a. Combining outcomes

In making predictions, one has to specify what the outcomes are. The issue is whether we should consider two or more outcomes as different or as variants of the same outcome. Sometimes this issue involves cases of abstractness. For instance, in the Latin negative prefix *in-*, used in English, there are several variants that show up: *il-* for words beginning with *l* (such as *illegal*), *ir-* for words beginning with *r* (such as *irregular*), and *im-* for words beginning with labials (such as *impossible*). When trying to predict the negative adjectival prefix in English, do we consider these four variants as a single morphological outcome (say, the abstract *IN-*) or as four different ones (*in-*, *il-*, *ir-*, or *im-*)? In general, our decision will affect our predictions of the negative adjectival prefix, and from those results we can perhaps discover which treatment (one or four outcomes) best accounts for speakers' actual predictions. For further discussion, see Chapman and Skousen 2005:12.

Another example of this problem of outcome specification arises in the case of the Finnish illative ending $-\mathrm{hV}_i\mathrm{n}$ (discussed in the section 4b). There we considered this ending as a single outcome, but theoretically one could consider it as a multitude of distinct outcomes, each different with respect to the copied vowel V_i:

voi ‘butter’	voihin	–hin
syy ‘reason’	syyhyn	–hyn
kuu ‘moon’	kuuhun	–hun
tie ‘road’	tiehen	–hen
työ ‘work’	työhön	–hön
suo ‘swamp’	suohon	–hon
pää ‘head’	päähän	–hän
maa ‘land’	maahan	–han

Again, the issue is empirical; and the best predictions occur if we treat all of these forms as the same outcome, not as eight distinct outcomes (the latter leads to an substantial increase in the heterogeneity of the contextual space and subsequent loss in predictability). Such an analysis argues that speakers are therefore aware of the basic identity of all these variant forms.

5b. Separating or combining the outcomes

Another issue deals with whether we have a single outcome or separate outcomes that apply in some order with respect to each other (or perhaps independently of each other). As an example of this, consider plural formation in German. The plural form can be viewed as two processes, adding an ending and mutating the stressed stem vowel (umlauting):

singular	*plural*	*ending*	*umlauting*
Berater ‘advisor’	Berater	Ø	no
Vater ‘father’	Väter	Ø	yes
Bauer ‘farmer’	Bauern	n	no
Motor ‘motor’	Motoren	en	no
Tag ‘day’	Tage	e	no
Band ‘volume’	Bände	e	yes
Band ‘ribbon’	Bänder	er	yes
Band ‘bond’	Bande	e	no
Band ‘band’	Bands	s	no

One issue here is whether *-n* and *-en* should be considered syllabic variants of the same ending. Another issue involves the case when the stressed stem vowel is already a front vowel; in that case, we may ask whether one should consider umlauting as vacuously applying or not at all:

singular	*plural*	*ending*	*umlauting*
Rücken ‘back’	Rücken	Ø	yes or no?
Bild ‘picture’	Bilder	er	yes or no?
Bär ‘bear’	Bären	en	yes or no?
Brief ‘letter’	Briefe	e	yes or no?

Ultimately, the issue is how tightly linked are the endings with the umlauting. For some endings (such as *-er*), we expect umlauting (whenever it can apply). For other endings (such as *-en* or *-s*), we do not expect umlauting (whenever it can apply). And for some endings (such as *-e*) we can have umlauting or not, depending on the word (and again, whenever it can apply). These links between the ending and umlauting suggest that we should consider the cases of plural formation as single outcomes. But ultimately, the issue is empirical. For instance, when the stressed stem vowel is not already a front vowel, do speakers (in the historical or dialectal development of the language or as children learning the language) remove the umlauting for the *-er* ending (which expects umlauting whenever it can apply)? If so, then we may wish to predict the ending and the umlauting separately from one another – or perhaps sequentially, with one being predicted first, then the other being predicted on the basis on the first prediction.

This example brings up the more paramount question of sequential versus simultaneous prediction in dealing with syntactic prediction and, we should add, virtually every other kind of linguistic prediction. Language processing involves sequencing through time, with one prediction following another and typically depending on previous decisions.

6. Repetition in the Dataset

The final issue that I would like to bring up here is the question of how exemplars should be represented in the dataset. In Skousen 1989, I almost always listed the exemplars for morphological problems as types rather than as tokens. And in most instances, types have worked much better than tokens in predicting morphological behavior. When tokens are specified, the highly frequently occurring types typically overwhelm the analysis. In Skousen 1989:54, I discuss the issue of types versus tokens and observe that "ultimately, the difference between type and token can be eliminated by specifying enough variables. By increasing the number of variables every token occurrence will also represent a single type." But whether this proposal is feasible is questionable since there is undoubtedly some empirical limitation on the number of variables that can be handled.

The need to distinguish between types and tokens in phonetic and morphological problems has been emphasized in Bybee 2001:96-136. Baayen and his colleagues (see de Jong, Schreuder, and Baayen 2000) have been arguing that a more accurate exemplar basis would be family types, where datasets would list all the morphologically related types, both inflectional and derivational, in datasets. Again, decisions of this sort regarding what to put in the dataset is an empirical issue.

7. References

Bybee, Joan (2001). *Phonology and Language Use.* Cambridge: Cambridge University Press.

Chapman, Don, and Royal Skousen (2005). "Analogical Modeling and Morphological Change: The Case of the Adjectival Negative Prefix in English". *English Language and Linguistics* 9:2, 1-25.

de Jong, Nivja H., Robert Schreuder, and R. Harald Baayen (2000). "The Morphological Family Size Effect and Morphology". *Language and Cognitive Processes* 15, 329-365.

Elzinga, Dirk (2006). "English Adjective Comparison and Analogy". *Lingua* 116, 757-770.

Hey, Anthony J. G., editor (1999). *Feynman and Computation: Exploring the Limits of Computers*. Reading, Massachusetts: Perseus Books.

Skousen, Royal (1989). *Analogical Modeling of Language*. Dordrecht, The Netherlands: Kluwer.

Skousen, Royal (1992). *Analogy and Structure*. Dordrecht, The Netherlands: Kluwer.

Skousen, Royal, Deryle Lonsdale, and Dilworth B. Parkinson, editors (2002). *Analogical Modeling: An Exemplar-Based Approach to Language*. Amsterdam: John Benjamins.

Skousen, Royal (2003). "Analogical Modeling: Exemplars, Rules, and Quantum Computing". *Proceedings of the Twenty-Ninth Annual Meeting of the Berkeley Linguistics Society*, edited by Pawel Nowak, Corey Yoquelet, and David Mortensen, 425-439 [also available at <http://humanities.byu.edu/am/>].

Skousen, Royal (2005). "Quantum Analogical Modeling: A General Quantum Computing Algorithm for Predicting Language Behavior". Preprint, posted under Quantum Physics on <arXiv.org>, quant-ph/0510146, 18 October 2005.

8. Acknowledgments

I wish to thank members of the Analogical Modeling Research Group at Brigham Young University for their helpful criticisms and suggestions for improvement: Deryle Lonsdale, David Eddington, Dirk Elzinga, Dana Bourgerie, and Don Chapman. I also wish to thank Benjamin Skousen for his comments on the Chinese classifiers.