Exemplar-Driven Analogy in Optimality Theory

James Myers
Graduate Institute of Linguistics
National Chung Cheng University
Min-Hsiung, Chia-Yi 621
Taiwan
lngmyers@ccunix.ccu.edu.tw

Presented at the Conference on Analogical Modeling of Language
Brigham Young University, Provo, Utah, March 23-24, 2000

[Slightly revised version to appear in R. Skousen, D. Lonsdale, & D. B. Parkinson (eds.)
Analogical Modeling: An Exemplar-Based Approach to Language. John Benjamins.]

1. Introduction

The term "analogy" may be something of a dirty word for most theoretical linguists, but it shouldn't be forgotten that it was theoretical linguists who first coined the term as it applies to language. Of course, when the Neogrammarians wrote about paradigm leveling or four-part proportional analogy, it was often just in passing on the way to what really interested them, namely regular rules. The same has been true for their structuralist and generative descendents, with a major excuse usually being that analogy was too vague a notion to deal with in a formal model. While this excuse is no longer valid, a sharp divide nevertheless remains between the mostly positive attitude of computer modelers and psycholinguists towards analogical approaches, and the mostly negative attitude of generative linguists.

Recently this has begun to change. Optimality Theory (or OT, to use its standard abbreviation; Prince and Smolensky 1993, 1997) is a formal generative model of language that has certain properties that make it capable of handling true exemplar-driven analogy (as opposed to earlier generative reanalyses of analogy using general rules, e.g. Kiparsky 1978, 1988). Recognition of this fact is gradually filtering through the "mainstream" OT literature, with prominent researchers such as Kenstowicz (1995, 1997), Steriade (1999a, 1999b, 2000), Burzio (1997a,b, 1999, 2000, to appear), and Hayes (1999b) beginning to peek out of the analogical closet, along with newer scholars trained in the generative tradition, including Benua (1995, 1997a,b), Alderete (1999), Kirchner (1999), and Albright (to appear). My own contribution has been to try to push the analogical approach as far as it can go in an OT formalism, in the hope that generative linguists and nongenerativists working on analogy can better share insights.

To this end, I set myself the goal of building a completely explicit formal model of the traditional linguistic notion of four-part proportional analogy (focussing on phonological analogy), using nothing but devices already found in the OT literature. In this paper I first review some of this literature to show that my model is not extremely radical by current generative standards, and then I describe how my model actually works. Next I prove that it is equivalent to the simplest possible kind of connectionist network, a linear associator, which has well-known strengths and weaknesses (see e.g. Anderson 1995). I then compare the explicit quantitative predictions made by the OT model with those made by Analogical Modeling of Language (AML; Skousen 1989, 1992). In general, the OT model of analogy performs much

worse than AML, but the fact that it makes quantitative predictions at all, and that these predictions are far more accurate than chance, convinces me that it is in principle possible to build a bridge between generative and nongenerative approaches to analogy. Moreover, I show how insights from AML and connectionism may be used to improve the quantitative accuracy of the model (though this requires going beyond OT formalism). Finally, like traditional generative theories of language, and unlike AML, my OT model represents both inputs and outputs with features, and it is also capable of incorporating nonanalogical factors. These two properties seem to give it an advantage in handling certain empirical phenomena, and so I hope that in building the bridge between OT and AML, the exchange of insights will run both ways.

2. Analogy in Optimality Theory

In this section I discuss some of the properties that make OT more similar to analogical approaches than previous generative models and show how explicitly analogical analyses are becoming more common in the OT literature. This discussion will then lead to the fully analogical OT model described in the following section.

The most obvious property that makes OT analogy-friendly is that it is non-derivational and surface-based. This results from its being a descendent both of standard generative theories of linguistic constraints and of so-called constraint-satisfaction connectionist networks (see especially Prince and Smolensky 1997).

Another important property of OT is that it makes a foundational distinction between two kinds of linguistic constraints. So-called Structure constraints include those like the famous NOCODA, which require output forms to conform to universal structural principles that may or may not be motivated by extra-linguistic factors. Such constraints are the clear descendents of generations of generative constraints, including the syntactic principles of Government and Binding theory. However, OT also posits so-called Faithfulness constraints, whose sole job it is to require forms to be "faithful" to themselves or to other forms, that is, to prevent the Structure constraints from doing anything. If OT grammars had only Structure constraints, all languages would be reduced to the maximally unmarked form, which is clearly not the case. With Faithfulness constraints, OT thus makes it explicit, perhaps for the first time in generative linguistics, that at least half of the human language faculty involves brute memorization of forms as they are, regardless of how inelegant, costly, or marked they may be. A purely analogical OT model, then, would be one that is built solely out of Faithfulness constraints.

The fact that exemplar-driven analogy is driven by exemplars may seem to pose an impossible challenge for OT, since OT constraints are usually described as completely general, even universal or innate. From the very beginning, however, it has been recognized that it is often necessary to posit constraints that are specific to specific classes, or even to particular lexical items. For example, in perhaps the most famous application of OT, McCarthy and Prince's (1993a) analysis of Tagalog *um* infixation, just such a constraint plays a crucial role. The claim of this analysis is that the distribution of *um* can be explained if one thinks of it as being affixed as close to the beginning of the word as possible without creating a new syllable coda. Disallowing the coda is the responsibility of the universal Structure Constraint NOCODA, but clearly there is no universal principle requiring affixes to appear towards the beginning of a word. To account for this fact, McCarthy and Prince (1993a) propose a universal Faithfulness constraint EDGEMOST which is parameterized by word edge (in this case, the left one) and by morpheme (in this case, *um*). Hence EDGEMOST(Left, *um*) requires the morpheme *um* to appear at the left edge of a word, meaning that the further away *um* is from this edge, the more it violates

this constraint. McCarthy and Prince (1993b) later reanalyzed EDGEMOST(Left, *um*) within the Generalized Alignment approach, now calling it ALIGN(*um*, Left, Stem, Left), but it still has to refer specifically to the morpheme *um*.

Universal constraints parameterized by lexical item are sometimes called parochial constraints (e.g. Hammond 1995), and they are ubiquitous in the OT literature. For example, to deal with the different phonological behaviors of the two major classes of English derivational morphology (e.g. the ∅~[n] alternation in *condemn-condemnation* vs. no alternation in *condemn-condemnable*), Benua (1997a,b) uses parochial constraints parameterized to each class, which then allows her to rank the constraints separately and derive the phonological differences (this analysis will be described in more detail below). Some OT researchers (e.g. Russell 1995, 1999, Hammond 1995, 1997, Golston 1996) have gone much further, proposing models in which morphemes or words are themselves (sets of) parochial constraints.

Thus to make analogy exemplar-driven in an OT model, we need parochial Faithfulness constraints. For the purposes of my model of analogy described in the next section, I maintain the standard OT assumption that distinguishes inputs (roughly equivalent to the underlying representations of earlier generative theories of phonology) from outputs (i.e. surface forms), and the familiar set of Faithfulness constraints called IDENT-IO which require input (I) and output (O) forms to be identical in some feature (McCarthy and Prince 1995). Somewhat new is my assumption that IDENT-IO is parochial rather than general, and that it operates over whole words, not individual morphemes. The general form of this parameterized constraint is IDENT-IO(*W*;F), where *W* represents a word, and F a feature. For example, the constraint IDENT-IO(*bat*;[labial]) would mean that the word *bat* cannot change its value of [labial] from input to output. Translated into more theory-neutral terminology, this kind of constraint has the job of preventing analogy (or other factors) from affecting one particular phonological property in one particular word.

The use of the parameters *W* and F require some brief comments. As is the case for any model of analogy, the particular representation used may have enormous consequences for how it works (see e.g. Baayen's 1995 comments on Skousen 1992, or Pinker and Prince's 1988 criticisms of Rumelhart and McClelland 1986). By calling F a "feature" I don't necessarily adopt the standard distinctive features of generative phonology; setting F to /b/ or VOT=20 msec or even [bæt] may prove to work better. Likewise, I don't necessarily follow the linguist's traditional focus on types rather than tokens. *W* thus may be taken to represent a particular token of a word (as spoken or heard by some individual). Token-based approaches to phonology are becoming more common (see e.g. Bybee 2000, Kirchner 1999), and I will also adopt this assumption in this paper, since as we will see, it allows my OT model to handle lexical frequency effects in a natural way.

Nevertheless, IDENT-IO is not the sort of Faithfulness constraint that can itself give rise to analogy, which of course involves relations *between* words. Fortunately, here is where recent developments in OT theorizing become particularly useful for analogical purposes. Starting with McCarthy and Prince (1995), Faithfulness has been generalized from involving only inputs and outputs, to involving any pairs of representations. McCarthy and Prince (1995) applied this new theory (called correspondence theory) to two parts of a single output (stem and reduplicant in reduplicated forms), and soon thereafter Kenstowicz (1995, 1997) and Benua (1995, 1997a,b) applied it to pairs of morphologically related output forms.

Output-output (OO) correspondence allows for analyses that are strikingly different from anything that had previously been allowed in generative theory, and strikingly similar to

traditional theories of analogy. For example, a blatant use of paradigm leveling forms the basis of Benua's (1997a,b) analysis of *condemn-condemnation/condemnable* alluded to earlier. In essence, her analysis suggests that while *condemn* may lose its (supposedly) underlying /n/ due to a Structure constraint against syllable-final [mn] sequences, the loss of the /n/ in *condemnable* is by analogy: a parochial Faithfulness constraint, specific to the class of morphology that includes *-able*, requires *condemn* and *condemnable* to share the property of [n]-lessness. Technically this is handled by ranking the anti-[mn] constraint at the top (it is never violated), then ranking the OO-constraint above the IO-constraint (i.e. it's better for *condemnable* to become similar to *condemn* than to keep its underlying /n/). Another parochial OO-constraint for morphology like *-ation* is ranked below the IO-constraint (i.e. it's better for *condemnation* to keep its original /n/ than to become similar to *condemn*).

Without necessarily condoning the particular application of analogy here, it's worth noting the important sociological development that such analyses represent. First, while there are some grumblings about them in the OT literature (e.g. Booij 1997, Hale, Kissock and Reiss 1998), they are becoming more common; other examples include Burzio (1997a, b, 1999, 2000, to appear) and Steriade (2000). Second, these authors openly acknowledge that what they are doing should be called analogy; Kenstowicz (1995, 1997) makes this particularly explicit. Third, analogical analyses of this sort have been accepted so rapidly that one has to conclude that they are filling a need that has long been felt but could never before be expressed.

For example, the standard generative phonology textbook Kenstowicz (1994) (written just before OT came to dominate phonological theory) argues that the vowel-length differences many speakers show before the flaps in *writer* and *rider* must be due to ordered rules (i.e. vowel-lengthening before flapping), just as argued in Chomsky and Halle (1968). Ironically, Kenstowicz (1994:71-72) does consider an alternative analysis in which *writer* contains a short vowel by analogy with *write* (more precisely, *writer* contains the short-vowel allomorph of *write*), but then rejects it. With output-output correspondence (developed partly with the help of Kenstowicz himself), the analogical analysis can now be formalized by positing a Faithfulness constraint IDENT-OO([vowel length]) that outranks the Structure constraints requiring vowels to be long before voiced consonants. Although I don't know of any work in the OT literature that actually presents this analysis, it's not difficult to flesh out the details. To illustrate this, and to give readers less familiar with OT notation a chance to practice before things get more technical later on, I provide the details here.

First a Structure constraint requiring consonants to be flapped in certain intervocalic environments (call it FLAP) must be ranked higher than the Faithfulness constraint IDENT-OO([vowel length]), which is in turn ranked above the Structure constraint requiring long vowels before voiced consonants (call it LONG). The following tableaux (as they are called) then illustrate what happens in the pairs *ride-rider* and *write-writer*. As is usual in the OT literature, I list possible outputs in the first column (here, all possible combinations of vowel length with flapping). Constraints are listed left to right from highest to lowest rank (an OT grammar is defined by a constraint ranking). Stars indicate violations of a given candidate output by a given constraint; multiple stars mean multiple violations by the same candidate. The optimal candidate (i.e. the one predicted to be grammatical, marked with a pointing finger) is in the subset of candidates that least violate the highest-ranked constraint, and in this subset, it is in the subset of candidates that least violate the second-highest-ranked constraint, and so on. Perhaps a quicker way to spot the optimal candidate is to mentally translate the stars into digits (* = 1, ** = 2, " " = 0, etc), and the row of stars for a given candidate into a number (e.g. 1020 for the first row in (1)). The optimal candidate is then the output associated with the lowest number (e.g. in

(1), the candidate marked with the pointing finger is associated with the number 0002).

Note that in (1), no analogy occurs. The optimal candidate here is simply the one that obeys both Structure constraints (FLAP and LONG). By contrast, in (2), the structurally best candidate is *rait-rai:Dr* (second from bottom), but that is not the one chosen. Instead the optimal candidate is one in which *write* and *writer* have vowels with the same duration, since IDENT-OO([vowel length]) outranks LONG.

(1) Why both *ride* and *rider* have long vowels (no analogy)

Input: raid-raidr	FLAP	IDENT-OO ([vowel length])	LONG	IDENT-IO ([vowel length])
raid-raidr	*		**	
ra:id-raidr	*	*	*	*
raid-ra:idr	*	*	*	*
ra:id-ra:idr	*			**
raid-raiDr			**	
ra:id-raiDr		*	*	*
raid-ra:iDr		*	*	*
☞ ra:id-ra:iDr				**

(2) Why both *write* and *writer* have short vowels (paradigmatic leveling)

Input: rait-raitr	FLAP	IDENT-OO ([vowel length])	LONG	IDENT-IO ([vowel length])
rait-raitr	*			
ra:it-raitr	*	*		*
rait-ra:itr	*	*	*	*
ra:it-ra:itr	*		**	**
☞ rait-raiDr			*	
ra:it-raiDr		*	**	*
rait-ra:iDr		*		*
ra:it-ra:iDr			*	**

The main point to take away is that by using output-output correspondence, the traditional generative rule-ordering analysis can be replaced with an analogical one (specifically, paradigmatic leveling), and this analogical analysis is formally precise. My proposed OT model of analogy, however, goes much further than the examples just sketched.

3. Four-part proportional analogy in Optimality Theory

To the best of my knowledge, nothing in the OT literature has taken the logical next step, which is to try to build an OT model of four-part proportional analogy. This more general form of analogy subsumes paradigm leveling as a special case, and it is far more powerful. Moreover, as I noted in the introduction, it is something with a long tradition in linguistics, and thus I hope less threatening to unconditioned generative linguists than more sophisticated models of analogy

like AML. In this section I show how to bring this kind of analogy into OT, focussing on technical issues (see Myers 2000a for discussion of the applications of the model to linguistic data that pose serious problems for traditional generative models without analogy; also Green 2001).

The first thing to do, it should be clear, is to make output-output correspondence completely parochial, rather than requiring that it only apply within paradigms. Otherwise we can't describe the irregularization of *dive* (past tense *dove*) as analogy with *drive-drove*. Thus I posit OO-constraints of the form IDENT-OO(W_i,W_j;F_k), where W_i and W_j are words (or word tokens) and F_k is some feature (in the extended sense of "feature" discussed earlier).

But of course analogy does not work to make any random pair of words similar to each other. To constrain the IDENT-OO constraints, we have to go somewhat beyond the OT mainstream, but only somewhat. The problem is this. In a proportional analogy, there are four items (*a*, *b*, *c*, *d*), standing in the relation *a*:*b*::*c*:*d*. This is standardly taken to mean that if *a* shares feature F with *c*, then *b* shares feature G with *d*. In terms of parochial IDENT-OO constraints, this says: if IDENT-OO(*a*,*c*;F) then IDENT-OO(*b*,*d*;G). Is there any way of creating a new constraint that is violated if and only if this logical implication is false?

As it happens, there is. In the grab bag of OT innovations is the notion of constraint conjunction, which creates new constraints with Boolean operators (see Smolensky 1995, Crowhurst and Hewitt 1997, and Balari, Marín, and Vallverdú 2000 for nonanalogical applications). It turns out that the constraint we need has the form given below (conjoined with the AND operator), which is violated if and only if at least one of the two component constraints is violated.

(3) IDENT-OO(*a*,*c*;F) ∧ IDENT-OO(*b*,*d*;G) [abbreviation: OO∧OO-(*a*,*c*;F)(*b*,*d*;G)]

That this constraint has the desired behavior can be seen if we consider a toy lexicon containing four items *a*, *b*, *c*, *d*. If *a* and *c* are already similar, as in (4a), *d* will change its form to conform to *b*. However, if *a* and *c* aren't already similar, as in (4b), *d* won't change (the conjoined constraint is violated in both candidates since the first component constraint is violated, and hence it has no effect on the choice of optimal output).

(4) a. *a* and *c* are similar

***a*=[+F]**, *b*=[+G], ***c*=[+F]**, *d*=[-G]	OO∧OO-(*a*,*c*;F)(*b*,*d*;G)	IO-(*d*;G)
d=[-G]	*	
☞ *d*=[+G]		*

b. *a* and *c* are not similar

***a*=[-F]**, *b*=[+G], ***c*=[+F]**, *d*=[-G]	OO∧OO-(*a*,*c*;F)(*b*,*d*;G)	IO-(*d*;G)
☞ *d*=[-G]	*	
d=[+G]	*	*

While this makes analogical change in *d* contingent on the properties of *a*, *b*, and *c*, there is still nothing preventing us from bringing a random quartet of words together into a spurious proportion. To deal with this, I fall back on the time-honored generative tradition of positing a

universal principle. This principle also explicitly disallows IDENT-OO constraints acting on their own outside of proportions.

(5) PROPORTION PRINCIPLE
Given the items *a*, *b*, *c*, *d* in a language and the features F and G, the conjoined constraint IDENT-OO(*a*,*c*;F)∧IDENT-OO(*b*,*d*;G) is generated if and only if there exists a single outcome function *o* such that *o*(*a*) = *b* and *o*(*c*) = *d*. IDENT-OO constraints do not exist outside of such conjoined constraints.

In justification of this move, I point out that all other models of analogy (including traditional notions, AML, and connectionism) tacitly assume something very much like this principle. For example, if one runs an AML simulation on data points associated randomly with outcomes (e.g. *drive-ate*, *strive-banana*), one shouldn't expect to get particularly insightful results. Any theory of analogy thus presupposes a theory of "relatedness"; the Proportion Principle merely makes this presupposition explicit.

This completes the set of supplemental devices needed for the OT model of analogy. For the remainder of its powers the model relies on nothing more than the central OT notion of extrinsic constraint ranking. This is all that is needed to deal with the notoriously capricious nature of analogy (which often fails to apply in one language in precisely the environment where it readily applies in another). For example, we can assume that all English dialects have constraints like the following, which requires the past tense forms of *drive* and *dive* to have the same vowel since the present tense forms have the same rime.

(6) IDENT-OO(*drive*,*dive*;[ayv]) ∧ IDENT-OO($drive_{PAST}$,$dive_{PAST}$;[o])

In a dialect where *dive* is regular, this constraint (whose existence is required by the Proportion Principle) is stripped of all power by being extrinsically ranked below the IO-constraint that keeps the past tense form of *dive* in its original form, as in (7a). By contrast, in a dialect where *dive* is irregular, these constraints are ranked in the reverse order, as in (7b), and the past tense of *dive* becomes *dove* by analogy with *drive-drove*.

(7) a. A *dive-dived* dialect

[drayv], [drov], [dayv], [dayvd]	IO-($dive_{PAST}$;[ay])	OO∧OO-(*drive*,*dive*;[ayv]) ($drive_{PAST}$,$dive_{PAST}$;[o])
☞ [dayvd]		*
[dov]	*	

b. A *dive-dove* dialect

[drayv], [drov], [dayv], [dayvd]	OO∧OO-(*drive*,*dive*;[ayv]) ($drive_{PAST}$,$dive_{PAST}$;[o])	IO-($dive_{PAST}$;[ay])
[dayvd]	*	
☞ [dov]		*

Paradoxically, extrinsic constraint ranking also turns out to provide a neat account of universal properties of analogy, such as gradient similarity effects, gang effects, and frequency

effects. The explanation for this is that under the null hypothesis, OT constraints can be extrinsically ranked in every possible way cross-linguistically. If we examine the quantitative predictions of the completely random ranking of analogical conjoined constraints, the probability that a given form will be changed by a given analogy is determined entirely by the number of triggering analogical constraints. For example, the more similar a target form is to an analogical trigger, the more features they will share, and thus the more analogical constraints there will be that are parochial with respect to those words (i.e. one such constraint per shared feature). Likewise, the larger the gang of analogical triggers $\{W_1, ..., W_n\}$ that are similar to a given target form W_{n+1}, the more analogical constraints there will be that are parochial with respect to those words (namely constraints referring to W_1 and W_{n+1}, W_2 and W_{n+1}, and so on).

The same argument works for frequency effects. To make this completely explicit, consider the following toy lexicon containing three word types *a*, *b*, *c*, where *a* and *b* are both equally similar to *c*, but *a* is twice as frequent as *b*. The question concerns which result the OT model predicts to be more likely: that *c* (in its form for *o*(*c*)) will analogize to *a* or that it will analogize to *b*.

(8) Lexicon: *a* = [+F], *o*(*a*) = [+G],
b = [+F], *o*(*b*) = [-G],
c = [+F]
Data set: {*a*, *a*, *b*}

Using constraints that are parochial with respect to tokens rather than types, the Proportion Principle generates the analogies given in the following tableau (analogies between *a* and *b* are left out, since we're focusing on the behavior of *c*). Note that there are two constraints enforcing similarity between *a* and *c*, and only one enforcing similarity between *b* and *c*. Note also that there is no claim that these constraints are extrinsically ranked in any particular way; following the convention in the OT literature, I indicate the lack of ranking by separating the constraint columns with dotted lines.

(9)

a = [+F], *o*(*a*) = [+G], *b* = [+F], *o*(*b*) = [-G], *c* = [+F]	OO∧OO- (*a*,*c*;F) (*o*(*a*),*o*(*c*);G)	OO∧OO- (*a*,*c*;F) (*o*(*a*),*o*(*c*);G)	OO∧OO- (*b*,*c*;F) (*o*(*b*),*o*(*c*);G)
o(*c*)=[+G]			*
o(*c*)=[-G]	*	*	

The question then becomes a mathematical one: given completely random constraint ranking, what is the probability that the candidate output *o*(*c*)=[+G] will be chosen as optimal? While the analogical flavor of this question is new, the issue of variable constraint ranking in OT is not. Going back to Kiparsky (1993), OT researchers have used variable ranking to deal with variable linguistic phenomena. Other applications of variable constraint ranking in OT include Anttila (1997), Anttila and Cho (1998), Nagy and Reynolds (1997), Hayes and MacEachern (1998), Boersma (1998), Boersma and Hayes (2001), and Myers (2000b). Most useful for our purposes here, Myers (2000b) proves several theorems for calculating precise probabilities without having to face the factorial explosion that occurs when all *n*! rankings of *n* constraints are

examined. The central result is what Myers (2000b) calls Anttila's Theorem (after Anttila 1997), stated below.

(10) ANTTILA'S THEOREM

If there are only two competing candidates X_1 and X_2, the probability that candidate X_1 will be chosen as optimal under completely random constraint ranking is

$P(X_1) = |C_{X1}| / [\ |C_{X1}| + |C_{X2}|\ ]$,

where $|C_{Xi}|$ = number of constraints that evaluate X_i over the alternative candidate

In other words, if there are only two candidates to consider, the probability that one will be optimal is just the proportion of constraints that favor it out of all constraints that favor either candidate (constraints that treat all candidates the same way can be entirely ignored, according to a theorem that Myers (2000b) calls Noncommittal Constraint Irrelevance).

Specifically, what we find with the analysis in (9) are the following probabilities: P(*o*(*c*)=[+G]) = 2/3, P(*o*(*c*)=[-G]) = 1/3. (Readers wanting to get a hands-on feel for Anttila's Theorem may write out all six (=3!) constraint rankings implied by the tableau in (9) to confirm that it does indeed work.) Thus *c* is twice as likely to conform to the analogy with *a* as with *b*. This demonstrates that this OT model shows one major kind of frequency effect: the more frequent the analogical trigger, the stronger its analogical force.

The model is also capable of handling the flip side of frequency effects, namely that the more frequent a potential analogical target, the less likely it is to undergo analogy (e.g. the blocking of regularization in high-frequency English verbs). To represent target frequency, we use token-parameterized IDENT-IO constraints. Continuing with the above example, we give word *o*(*c*) an initial value of [-G] and a token frequency of 2, resulting in the following tableau (the first three constraints are the same as in (9)). Anttila's Theorem now predicts the probabilities P(*o*(*c*)=[+G]) = 2/5, P(*o*(*c*)=[-G]) = 3/5. Thus an increase in the frequency of an analogical target decreases its likelihood of undergoing an analogy (here, a drop in P([+G]) from 0.667 to 0.400).

(11)

a = [+F], *o*(*a*) = [+G], *b* = [+F], *o*(*b*) = [-G], *c* = [+F], *o*(*c*) = [-G]	OO∧OO- (*a*,*c*;F) (*o*(*a*),*o*(*c*);G)	OO∧OO- (*a*,*c*;F) (*o*(*a*),*o*(*c*);G)	OO∧OO- (*b*,*c*;F) (*o*(*b*),*o*(*c*);G)	IO-(*o*(*c*);G)	IO-(*o*(*c*);G)
o(*c*)=[+G]			*	*	*
o(*c*)=[-G]	*	*			

This, then, is an OT model of true exemplar-driven analogy. It assumes virtually nothing that has not already been discussed in the OT literature, and its major technical devices (output-output correspondence and constraint ranking) are entirely mainstream. To understand precisely where the OT model stands among nongenerative models of analogy, however, we need to examine the nature of its quantitative behavior more closely. This is the subject of the following section.

4. Analogy in Optimality Theory and connectionism

In this paper I have been using the term "analogy" to refer to an empirical fact that has been recognized by linguists for almost two hundred years, not just the particular theory of it provided by AML. Thus I have no qualms in listing connectionism as an alternative model of analogy. For example, Rumelhart and McClelland (1986), using a connectionist model, is one possible analogical analysis of English inflection; Derwing and Skousen (1994), using AML, is another. In this section I show that the OT model of analogy sits squarely in the connectionist tradition. In fact, under a reasonable representational assumption (also made in AML), it is exactly equivalent to the simplest kind of connectionist network, a linear associator. If a more complex representational scheme is used, its behavior is somewhat more complex, but still essentially connectionist-like.

The representational assumption just alluded to involves supposing that the outcomes (i.e. the forms that the function *o* maps to) are atomic units, rather than composite forms built out of the same features that compose the data points. AML makes this assumption quite clearly (as do nearest-neighbor approaches; see elsewhere in this volume). For example, in the toy example in Skousen (1989:23-37), the basic forms are built out of three four-valued features, giving representations like 310 and 032, but the outcomes are the two distinct atoms *e* and *r*. It is not even immediately obvious how AML could be modified so that the outcomes themselves could be built out of features in any meaningful way (though I make an explicit suggestion in this direction in a later section). The atomic nature of the outcomes in AML makes it eminently suitable for morphological analogy, which involves choosing among a fixed set of distinct morphemes, but it may cause problems for certain kinds of phonological analogy, which may affect only part of a form at a time. This possible weakness of AML will be discussed further below, but first I will adopt the atomicity assumption and see what consequences it has for the OT model of analogy.

The general situation is as follows. We have a set of words (or word tokens) W_1, ..., W_n, represented with features F_1, ..., F_m, and an outcome function *o* mapping the words onto a set of atomic outcomes X_1, ... , X_a. We want to know what analogy will do with a new word W_{n+1} given all possible rankings of all conjoined analogical constraints conforming to the Proportion Principle. How can we calculate the relative probabilities $P(o(W_{n+1})=X_1)$, ... , $P(o(W_{n+1})=X_a)$?

At first this may seem like a very difficult problem. Since more than two candidate outputs are being considered, Anttila's Theorem does not apply. Moreover, the behavior of the constraints may possibly vary quite unpredictably. This would leave us with the computationally irritating factorial problem of checking all possible constraint rankings. As it happens, however, the assumption of atomic outcomes makes the constraints so well behaved that a slight extension of Anttila's Theorem can be used.

First, we can completely ignore all constraints that make no reference to W_{n+1} (e.g. those that require identity between W_1 and W_2). These will be vacuously obeyed by all possible outputs for W_{n+1}, and as stated by the theorem of Noncommittal Constraint Irrelevance mentioned earlier, constraints that don't choose among any candidates can be removed without affecting probabilities under variable ranking. Now, all analogical constraints that do refer to W_{n+1} must have the following form if they are to conform to the Proportion Principle. Note that in accordance with the atomicity assumption, the outputs $o(W_i)$ and $o(W_{n+1})$ in the second component of the conjoined constraint are completely identical, rather than merely sharing the value for a single feature.

(12) IDENT-OO(W_i,W_{n+1};F_j) ∧ IDENT-OO($o(W_i)$,$o(W_{n+1})$)

Logically there are only four possible behaviors of this constraint. These are represented schematically in the following table, where the stars indicate under what conditions the constraint is violated.

(13)

	IDENT-OO(W_i,W_{n+1};F_j)	¬ IDENT-OO(W_i,W_{n+1};F_j)
IDENT-OO($o(W_i)$,$o(W_{n+1})$)		*
¬ IDENT-OO($o(W_i)$,$o(W_{n+1})$)	*	*

Since our candidate outputs consist solely of possible outcomes for W_{n+1}, without varying the representation of W_{n+1} itself, the component constraint IDENT-OO(W_i,W_{n+1};F_j) must be either always obeyed or always disobeyed (for any given *i* and *j*) across the whole set of candidate outputs. If it's disobeyed (i.e. W_i and W_{n+1} are not identical in feature F_j), then the conjoined constraint in (12) will evaluate all candidate outputs as a violation. In this case, Noncommittal Constraint Irrelevance means we can ignore this particular conjoined constraint. However, if this component constraint is obeyed (i.e. W_i and W_{n+1} are identical in feature F_j), then the final decision is left to the other half of the conjoined constraint, namely IDENT-OO($o(W_i)$,$o(W_{n+1})$).

Under what circumstances is IDENT-OO($o(W_i)$,$o(W_{n+1})$) obeyed? Here is where the assumption of atomicity is crucial. Given this assumption, this constraint is obeyed if and only if the outputs of W_i and W_{n+1} are entirely identical, which means there is some atomic outcome X_k such that $o(W_i) = o(W_{n+1}) = X_k$. This means that this constraint is violated (along with the entire conjoined constraint) whenever $o(W_{n+1})$ is some atomic outcome other than X_k. Thus if the premise is true (i.e. W_i and W_{n+1} are identical in some feature), the conjoined constraint will be violated by all candidate outputs except one (namely the one where $o(W_{n+1}) = X_k$).

For example, suppose that $o(W_1) = X_1$, and that W_1 and W_{n+1} are identical in feature F_1 (e.g. they both share value $[+F_1]$) but not in feature F_2. One corner of the resulting tableau will thus appear as follows, where the stars in the bottom row symbolize the consistent violation of these constraints for all candidate outputs other than $o(W_1) = X_1$.

(14)

	OO∧OO- (W_1, W_{n+1};F_1)($o(W_1)$, $o(W_{n+1})$)	OO∧OO- (W_1, W_{n+1};F_2)($o(W_1)$, $o(W_{n+1})$)	...
$o(W_{n+1}) = X_1$		*	...
$o(W_{n+1}) = X_2$	*	*	...
...	*	*	...

In general, then, these constraints only act in two ways: either they don't do anything, or they reject all candidate outputs but one. This limitation of winners to at most one per constraint makes a slightly modified version of Anttila's Theorem applicable. The proof is virtually the same as that given in Myers (2000b) for Anttilla's Theorem, and may be informally stated as follows. Given Noncommittal Constraint Irrelevance, we only have to consider constraints that pick a single winner. The probability that a given candidate will win overall, then, is simply the probability that a constraint that favors it is ranked at the top (thus making all the other constraints powerless).

(15) Anttila's Theorem for constraints that choose at most one winner

If all constraints evaluate at most one candidate as optimal, then the probability that candidate X_1 is optimal overall, given completely random constraint ranking, is

$P(X_1) = |C_{X1}| / [\ |C_{X1}| + |C_{X2}| \ldots + |C_{Xa}|\ ]$,

where $|C_{Xi}|$ = number of constraints that evaluate X_i over the alternative candidates

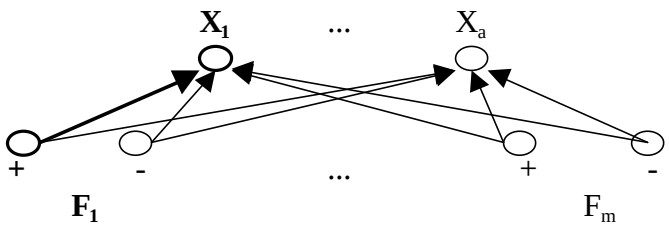
This theorem has a special interpretation in our case, however. The set C_{Xi} (i.e. the set of all constraints that evaluate X_i over the alternative candidates) contains all analogical constraints that require W_{n+1} to match some analogical trigger word W_k in some feature and that require the outcomes for W_{n+1} and W_k both to be X_i. Keep this sentence in mind; a variant of it will return shortly.

Now I proceed to show that this model works precisely the same way as a linear associator. In this simplest of all connectionist networks (see Anderson 1995 for a lucid introduction and references), there are two layers of nodes, and each node is connected to all other nodes in the other layer. In training such a model, two vectors of node activations are presented to the model, and learning occurs through a Hebbian rule, i.e. a connection is strengthened if the two nodes that it connects are simultaneously activated. In our case, node activations during training must be either 0 or 1, all connection weights are initialized to 0, and the rule increases a connection weight by adding 1 if and only if both connected nodes have activation 1. Using standard connectionist notation, the rule can be stated as follows.

(16) $\Delta w_{ij} = a_i a_j$,
where w_{ij} is the weight of the connection between nodes i and j; a_i and a_j are the activations of nodes i and j, respectively; and Δw_{ij} represents the amount added to w_{ij} each time a_i and a_j are changed.

As for the architecture of the network, one layer will of course consist of a set of nodes for the atomic outputs $X_1, \ldots, X_a$. The other must consist of sets of nodes representing values of the features $F_1, \ldots, F_m$, used for the word forms $W_1, \ldots, W_n$. For example, if these features were binary, the architecture would be as shown in the following diagram.

(17)



To see how the model works, consider the situation illustrated above in (14), where $o(W_1)=X_1$, and W_1 and W_{n+1} both share value $[+F_1]$. During training, W_1 (represented in features) would be presented to the bottom layer, thus activating the node $[+F_1]$, and $o(W_1)$ would be presented to the top layer, thus activating the node X_1. Since these two nodes would be

simultaneously activated, the Hebbian learning rule would add 1 to the weight of the connection between them (highlighted in the above diagram).

When training is complete, we freeze learning and present the word W_{n+1} (represented in features) to the bottom layer. What we want to know is the relative activation of the outcome nodes X_1, ... , X_a, since this indicates how likely it is that $o(W_{n+1})$ is to be realized as one of these outcomes. As in all connectionist networks, the activation for each outcome node is derived from the sum of the weights of the connections leading into it from active nodes. The most commonly used connectionist networks today (e.g. feedforward networks trained with backpropagation, Hopfield networks, and so on) take this input sum and then run it through a nonlinear function. What makes a linear associator linear is that it does not: the activation of an output node is directly proportional to the sum of the incoming weights. Our activation function is thus as follows.

(18) $a_i = \Sigma_j a_j w_{ij}$,
where a_i and a_j are the activations of nodes i and j, respectively, and w_{ij} is the weight of the connection from input node j to output node i.

Surprisingly, perhaps, this architecture and these equations mean that the activation of outcome node X_1, relative to all the other node activations, is calculated in precisely the same way as the probability $P(X_1)$ in the OT model (i.e. in (15)). Since in our example we know that W_{n+1} has feature value $[+F_1]$, the weight of the highlighted connection in (17) between $[+F_1]$ and X_1 will be added into the activation of node X_1. This weight itself represents the number of instances in which two things are simultaneously true: a word W_i has the value $[+F_1]$ and the outcome $o(W_i)$ of this word is X_1. In general, the activation of X_i will represent the total number of instances such that W_{n+1} matches some analogical trigger W_k word in some feature and such that the outcomes for W_{n+1} and W_k are both X_i. (Here is the reappearance of the sentence, in modified form, that I asked the reader to keep in mind earlier.) In other words, the activation of node X_i is precisely identical to $|C_{Xi}|$ (i.e. total number of X_i-favoring constraints in the OT model). This in turn means that the proportion of activations represented by X_i is given by the formula in (15), which divides $|C_{Xi}|$ by the sum of all "non-noncommittal" constraints (which is equivalent to the sum of all output activations in the network model). Under the atomicity assumption, then, the OT model of analogy is equivalent to a linear associator.

Linear associators have very well-understood strengths and weaknesses (see Anderson 1995 for discussion). Among their strengths is the fact that they are actually found in the nervous systems of some simple animals, and more to the point here, that they capture the essential properties needed for analogy (namely, the properties described in the previous section, such as gradient similarity effects, gang effects, and frequency effects). Among their weaknesses are technical mathematical limitations that may not be relevant here (e.g. they cannot distinguish nonorthogonal vectors in the training set, and like all two-layer networks they cannot learn exclusive-or or parity), but they also suffer from a problem that makes them less than ideal for the quantitative analysis of analogy: they are overly indecisive. Due to the lack of a nonlinear activation function, they tend to waver between states rather than showing crisp categorical behavior (categoricality is something of a problem for connectionism in general, but linear associators are absolutely abysmal).

Before demonstrating these strengths and weaknesses in the next section by pitting the OT model directly against AML, we should first briefly consider what happens if we discard the

assumption that outcomes must be atomic. Unlike (standard) AML, the OT model has no problem using featural representations for outcomes. Consider again the general situation, identical to the one we have been examining, but where the set of candidate outcomes consists of all possible representations generated by the features F_1, ..., F_m. The analogical constraints will then have the following form, where the second component constraint now refers to just one feature.

(19) IDENT-OO(W_i,W_{n+1};F_j) $\wedge$ IDENT-OO($o(W_i)$,$o(W_{n+1})$;F_k)

These constraints no longer choose at most one candidate as optimal. Instead, they either evaluate all candidates the same (namely, as in the previous discussion, if the words W_i and W_{n+1} don't match in feature F_j), or they accept some of the candidates (i.e. if $o(W_i)$ matches $o(W_{n+1})$ in feature F_k) and reject the rest. If such a constraint favors any candidates at all, the number of favored candidates will almost always be greater than one. For example, if there are *m* features all with the same valency *v*, then each non-noncommittal constraint will favor v^{m-1} candidates. The tableau below illustrates this with three binary features, where $v^{m-1} = 4$.

(20)

W_1 = [+F,+G,+H], W_{n+1} = [+F,+G,-H], $o(W_1)$ = [-F,-G-H]	OO$\wedge$OO-(W_1, W_{n+1};F)($o(W_1)$,$o(W_{n+1})$;G)	...
$o(W_{n+1})$ = [+F,+G,+H]	*	...
$o(W_{n+1})$ = [+F,+G,-H]	*	...
$o(W_{n+1})$ = [+F,-G,+H]		...
$o(W_{n+1})$ = [+F,-G,-H]		...
$o(W_{n+1})$ = [-F,+G,+H]	*	...
$o(W_{n+1})$ = [-F,+G,-H]	*	...
$o(W_{n+1})$ = [-F,-G,+H]		...
$o(W_{n+1})$ = [-F,-G,-H]		...

These considerations show that the versions of Anttila's Theorem we used earlier cannot apply here, since no single constraint can alone be responsible for choosing an output candidate as optimal. However, we still don't have to rank all the constraints every possible way and tally up the results, because a given outcome candidate $o(W_{n+1}) = [\alpha_1F_1,..., \alpha_mF_m]$ (where α_i represent feature values) can still only be chosen as optimal under very well-defined circumstances. Namely, in order for this candidate to be chosen, it must be that for every feature F_i, at least one constraint favoring $[\alpha_iF_i]$ must outrank all constraints that favor $[\neg\alpha_iF_i]$ (i.e. any other value for this feature); this follows directly from the definition of OT constraints and constraint ranking (see Samek-Lodovici and Prince 1999 for more on the foundational mathematics of OT). If $|C[\alpha_iF_i]|$ and $|C[\neg\alpha_iF_i]|$ represent, respectively, the number of constraints that favor $[\alpha_iF_i]$ and the number that favor $[\neg\alpha_iF_i]$, we can use the reasoning behind Anttila's Theorem to deduce that the probability that the optimal candidate contains $[\alpha_iF_i]$ must be as follows.

(21) $P(o(W_{n+1})=[...\alpha_iF_i...]) = |C[\alpha_iF_i]| / [\ |C[\alpha_iF_i]| + |C[\neg\alpha_iF_i]|\ ]$

Now, two constraints that refer to different features (i.e. a constraint that favors $[\alpha_i F_i]$ and a constraint that favors $[\alpha_j F_j]$) do not interact at all. That is, no matter how they are ranked with respect to each other, the outcome will be the same. In lieu of a formal proof, I offer the following tableau to ponder, where the relative ranking of the constraints *[+F] and *[-F] (and of *[+G] and *[-G]) does indeed affect which candidate will win, but not the relative ranking of *[+F] and *[+G] (nor of *[-F] and *[+G] and so forth). For example, if *[+F] is ranked above *[-F] as shown, the first two candidates can never win no matter how *[+G] and *[-G] are ranked, even if one or both outranks *[+F] (try it and see).

(22)

	*[+F]	*[-F]	*[+G]	*[-G]
[+F, +G]	*		*	
[+F, -G]	*			*
[-F, +G]		*	*	
[-F, -G]		*		*

What this means is that the probability that a given ranking chooses a candidate containing $[\alpha_i F_i]$ is independent of the probability that this optimal candidate contains $[\alpha_j F_j]$ as well. In the above tableau, for example, the probability that the optimal candidate contains [+F] is 1/2 (by the formula in (21), which can also be confirmed by hand), and the probability that it contains [+G] is also 1/2. Neither fact is dependent on the other in any way. This allows us to apply the multiplication rule from probability theory, deriving the probability $P([+F,+G]) = 1/2 \cdot 1/2 = 1/4$ (which you may confirm by examining all 24 possible rankings of the constraints in (22)). In general, the probability that $o(W_{n+1}) = [\alpha_1 F_1,..., \alpha_m F_m]$ is given by the following formula (completing the proof is left as an exercise for the reader).

(23) $P(o(W_{n+1}) = [\alpha_1 F_1,..., \alpha_m F_m]) = \Pi_i |C[\alpha_i F_i]| / [|C[\alpha_i F_i]| + |C[\neg\alpha_i F_i]|]$

Interpreting this in connectionist terms is more difficult than when we made the atomicity assumption, but it does seem to have some interesting properties. As before, the number $|C[\alpha_i F_i]|$ can be thought of as the sum of all the connection weights leading into an output node, this time representing the feature value $[\alpha_i F_i]$. Now, however, we have something like a nonlinear activation function, or more properly, a function that takes as arguments the activations of, for each feature, the feature value node of the target outcome relative to the activations of the nodes for the other values for that feature. This function may tend to make the model more decisive, since it involves multiplication rather than merely addition, but since the multiplication involves fractions less than or equal to 1, it can only work to decrease activation. Further thought is needed to explore the quantitative implications of this aspect of the model, and I won't discuss this further in this paper. The primary point to note here is that while representing outcomes with features is not possible in (standard) AML, it poses no special difficulty in the OT model of analogy (at least from the theoretical side).

5. Analogy in Optimality Theory and AML

Given the lack of sophistication of the OT model beneath all of its complex notation, one might expect it to perform rather poorly when confronted with actual analogical tasks to carry out.

In this section I show that it does indeed perform much worse than AML. Nevertheless, at a higher level of description, the OT model actually performs remarkably well given its generative origins: in virtually every case, it correctly chooses which of the alternative outcomes should be the preferred one. Its weakness lies solely in the degree of probability it assigns to this outcome (always much lower than it should). I end the section by suggesting how the quantitative predictions of the OT model might be improved by borrowing ideas from connectionism, and alternatively, how the model could be made into a notational variant of the AML algorithm, with interesting consequences for AML itself.

Consider first an AML analysis of Finnish past tense *ti* ~ *si* allomorphy (namely the one in Skousen 1992:310-322; more recent AML analyses of this and related problems in Finnish are found in Skousen 1989 and in this volume). In this analysis, attention was restricted to a small set of two-syllable verbs ending in tAA (where A represents a low vowel), some of which form the past tense with the *ti* allomorph, some with the *si* allomorph, and some with either (at particular token frequencies of occurrence that vary word by word). Unpacking the description in Skousen (1992), the analysis uses seven contextual variables (i.e. features), listed in the following figure (with my own labels).

(24) [±C1]: a binary feature representing the presence or absence of an initial consonant
[C1 value]: a multivalued feature representing the onset consonant or the lack thereof
[V1 value]: represents the first vowel
[±V2]: represents the presence or absence of a second vowel
[V2 value]: represents the second vowel or the absence thereof
[±C2]: represents the presence or absence of a stem-final consonant
[C2 value]: represents the stem-final consonant or lack thereof

Skousen (1992) first gives the model a data set of 42 verbs, and then tests it on verbs not in the data set, including *viertää* "to slope". AML predicts that the probability of choosing the allomorph *ti* for this verb is very low: P(*vierti*) = 0.00153. The model therefore both picks up on a real pattern in the data, and is very decisive about its response. The result is so sharp that it appears as if it's due to a rule. Based on the data given, it is tempting for a linguist (e.g. myself) to state such a rule, namely if the stem is a closed syllable, choose *si*. Nevertheless, the mechanism used here is actually analogy, not a general rule; AML even allows one to list forms by their degree of responsibility for the analogy. Skousen (1992:321) ends his discussion by pointing out that the three factors affecting the strength of the analogy here are (using my terminology) gradient similarity effects between analogical trigger and target, the frequency of the analogical trigger, and (using the original wording) the "extensiveness of the homogeneity".

As I've shown in previous sections, the OT model captures the first two of these three factors, but like connectionism and other non-AML models of analogy, it ignores homogeneity. How far can the OT model get with the same data set, the same features, and the same test word *viertää*? To find out, I calculated the predicted probabilities for the two allomorphs given all possible rankings of all possible analogical conjoined constraints conforming to the Proportion Principle (or equivalently, their relative degree of activation in a linear associator). In practical terms, what I did was as follows. For each data verb stem, I counted the number of feature values that matched those in the stem *vier*, multiplied this sum by the number of tokens of *ti* and *si* reported for this data verb, and finally added up the totals for *ti* and for *si*. Some of my calculations are shown in the following table, which also includes the grand totals of the

activations for *ti* and *si* given the input *vier*.

(25)

	a	b	c	d	e	f	g	h	i	j	k	l
vier	+C1	/v/	/i/	+V2	/e/	+C2	/r/	a+...+g	No. ti	No. si	h·i ***ti***	h·j ***si***
hoi	1	0	0	1	0	0	0	2	26	0	52	0
i	0	0	1	0	0	0	0	1	2	0	2	0
kiel	1	0	1	1	1	1	0	5	0	22	0	110
kier	1	0	1	1	1	1	1	6	0	16	0	96
...	...	...	...	...	...	...	...	...	...	...	...	...
										Total:	**954**	**1475**

The predicted probability is thus P(*vierti*) = 954/(954+1475) = 0.39275. The fact that this number is less than 0.5 means that the OT model agrees that the preferred form should actually be *viersi*, the correct outcome. That is, under a winner-take-all interpretation, the OT model and AML both choose the same outcome. However, the probability P(*vierti*) predicted by the OT model is of course far higher than the near-zero probability predicted by AML. This may be a consequence of the OT model's linear (i.e. indecisive) nature, and/or it may be related to its ignoring homogeneity.

Other differences in the behavior of the OT model and AML can be seen if we list the data verb stems in order by their relative contribution to the analogical effect. In the OT model, ranking is by the number of feature matches weighted by token frequency (i.e. the values listed in the last two columns in (25)). The following table lists the ten most influential items according to each model (data for AML is from Skousen 1992:320).

(26)

AML			OT		
Verb stem	Outcome	Effect	Verb stem	Outcome	Effect
kier	si	0.270	pi	*ti*	748
piir	si	0.258	tie	si	432
kiel	si	0.169	pyy	si	180
siir	si	0.086	piir	si	120
rien	si	0.061	löy	si	112
pyör	si	0.018	kiel	si	110
viil	si	0.009	ve	*ti*	100
mur	si	0.009	kier	si	96
kiil	si	0.005	myön	si	90
vään	si	0.003	kään	si	78

Examination of this table shows that whatever the ultimate cause, the proximal cause of the OT model's quantitative problems is that it is too easily fooled by false analogies. The stem *pi*, for example, has a large influence merely because it happens to share two features with *vier* (i.e. [+C1] and [V1=/i/]), and because it is high frequency. That's too bad for the model, since the outcome for *pi* is *ti*, not the desired *si*. In spite of such problems, the OT model does manage to

include in the top ten some of the items that AML also considers important, namely the stems *piir*, *kiel*, and *kier*. Moreover, two other items ranked highly by the OT model (but not by AML) have the "correct" syllable structure according to the linguistic analysis, namely the closed-syllable stems *myön* and *kään* (AML instead lists *rien*, *pyör*, *viil*, *mur* and *vään*). This means that fully half of the ten most influential data points for the OT model are precisely the ones that should have the most influence. Given the extreme simplicity of the OT model, this is a rather remarkable achievement.

The unimpressive level of quantitative accuracy leaves room for improvement, of course. One way to improve it is to hand-pick the features in accordance with the linguistic analysis (mentioned above) that states that the crucial factor is syllable structure alone. This means that we ignore all features except [±C1], [±V2] and [±C2]. Carrying out the procedure, we end up with the predicted probability P(*vierti*) = 0.07368, which is far closer to zero, as desired. We might be able to justify this move if it were a cross-linguistic universal that suffix allomorphy is always sensitive only to syllable structure, thus implying an innate cognitive bias for some features over others. No such universal exists, however, and this move should rightly be dismissed as outright cheating.

To help learn how the accuracy of the OT model may be improved in more appropriate ways, I considered another simple example comparing the OT model with AML (and also with standard connectionism). This is the toy problem described in Baayen (1995:395) (based on an example in Skousen 1992:266-272) in which there are 22 data points composed of three features (which I will label F_1, F_2, and F_3). Each feature can take one of four values (0, 1, 2, 3), so the data points can be represented as strings like 002 or 332. There are two possible outcomes (A or B) which are purposely chosen to conform to a simple rule: if $F_2 \in \{0,1\}$, then the outcome is A, otherwise it is B. The existence of this rule can be seen by the regular distribution of A's and B's in the following table.

(27)

$F_{1\rightarrow}$	0				1				2				3			
$F_{2\rightarrow}$ $F_{3\downarrow}$	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3
0			B	B		A	B			A		B	A			
1		A		B					A			B				B
2	A		B						A	A					B	B
3					A					A		B		A		

Given these three four-valued features, there are 42 other possible data points (the empty cells in figure (27)). AML does not give a strictly categorical response in most of these cases. For data point 000, for instance, it predicts the probability P(A) = 0.871, rather than P(A) = 1 as required by the rule. (Of course, this begs the question of how easily human beings could also see this particular pattern as rule-governed.) Nevertheless, AML is never wrong about which outcome should be more probable. More importantly, its error rate is very low. This can be calculated with a number of methods; I used two. In the first method, I took all the data points for which the rule predicts P(A) = 1 and subtracted from 1 the actual probability provided by AML (e.g. 0.871 in the above case). The mean error, calculated this way, was a quite respectably low 0.057 (chance performance of course would be 0.5). As another measure of error rate, for every test point I calculated the Euclidean distance (commonly used in studying

connectionist models) between AML's predicted values for P(A) and P(B) and the correct values. For example, for data point 000, the correct probabilities for A and B respectively are (1, 0). AML predicted (0.871, 0.129). The Euclidean distance between these two points is 0.182. Here chance performance would be half the length of the diagonal of a unit square (i.e. 0.707); AML's mean error value was the still very low 0.081.

How does the OT model fare on the same data? Again, it depends on how you look at it. The table below shows the OT model's predictions for preferred outcome for the 42 test points (the original data points are shaded). The model only made one mistake, incorrectly claiming that P(223A) = P(223B). Given not just the simplicity of the OT model but also the sparse and scattershot evidence for the AB rule, I suggest that this should count this as another success.

(28)

$F_{1\rightarrow}$	0				1				2				3			
$F_{2\rightarrow}$ / $F_{3\downarrow}$	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3
0	A	A	B	B	A	A	B	B	A	A	B	B	A	A	B	B
1	A	A	B	B	A	A	B	B	A	A	B	B	A	A	B	B
2	A	A	B	B	A	A	B	B	A	A	B	B	A	A	B	B
3	A	A	B	B	A	A	B	B	A	A	***A/B***	B	A	A	B	B

A closer look at the results reveals the usual quantitative problems, however. Using the same measures of accuracy applied to AML, the mean error rate shown by the OT model's predictions of P(A) was 0.343 (chance = 0.500), and the mean error rate by Euclidean distance was 0.486 (chance = 0.707). Both results are better than chance, but the error rate is still far higher than that for AML.

Earlier I mentioned two possible causes for the quantitative shortcomings of the OT model: its lack of a nonlinear activation function, and its ignoring of homogeneity. To better understand which is responsible in this case, I compared the behavior of the OT model with that of a two-layer connectionist network which does have a nonlinear activation function, but as with connectionism in general is not particularly sensitive to homogeneity. Like the linear associator associated with the OT model, the output layer of this connectionist network consisted of just two nodes (one for each outcome A and B) and three sets of four input nodes (for the three four-valued features). The difference was that the network used a sigmoid (i.e. S-shaped) activation function and was trained using the backpropagation learning algorithm; since it only had two layers, this made it essentially equivalent to a perceptron (again I recommend Anderson 1995 for lucid discussion of these concepts). This model (simulated using the **`tlearn`** software; see Plunkett and Elman 1997) had absolutely no trouble learning the AB rule. I assumed that given any particular input, the activation of an output node (always between 0 and 1) represented the degree of probability that the model assigned that outcome given that input (a commonly made interpretation in the connectionist literature). Its error rate for P(A) was 0.038 and by Euclidean distance 0.085, roughly as low as for AML. It appears, then, that at least for this particular simple problem, the accuracy of the OT model might be improved simply by giving it a nonlinear activation function.

How could this be accomplished? The simplest nonlinear activation function used in connectionist models is a step function. This is a function that has some constant value (say 0) for all inputs below some threshold, and some other constant value (say 1) for all inputs above

the threshold. Unfortunately, this idea is doomed from the start, since the output node activations will now always be just 0 or 1, which makes it impossible to interpret them as continuously varying probabilities.

The particular sigmoid function used with back-propagation and other connectionist models is unlikely to be coaxed from the simple mathematics underlying the OT model. Continuous nonlinear effects may arise if outcomes are represented with features, as discussed at the end of the previous section, but this can't help us with the Finnish and toy problems examined here, which use atomic outcomes. Moreover, any other attempt to create a continuous sigmoid activation function must face the problem of where to locate the flexion point (i.e. the threshold). In most connectionist models, this point is located where the input is 0, but this can't work for the OT model as it currently stands. In the linear associator associated with the model, the input activation nodes and the connection weights are always nonnegative, making it impossible to have a negative sum feeding into the output node activation function. Thus if we maintain the general structure of the OT model, the location of the threshold must somehow be made to depend on the size of the training set (since connection weights increase arithmetically as more items are trained).

An alternative way to derive a continuous nonlinear activation function might be to posit "evil twins" for the IDENT-OO constraints, that is constraints like DIFF-OO(W_i,W_j;F_k) that require words W_i and W_j to have *different* values for feature F_k. Although there is no precedent for such constraints in the OT literature, by including measures of difference we would make the model more consistent with theories of comparison in the cognitive science literature (e.g. Tversky 1977). It is also possible that DIFF-OO constraints could be made to interact with the IDENT-OO constraints in such a way as to create the equivalent of a sigmoid activation function. This is because there is only one way that two words W_i and W_j can be completely identical, and only one way they can be completely different (assuming binary features), but there are many ways that they can be partially similar and partially different. The resulting binomial distribution is an approximation of the normal distribution, which in turn approximates the first derivative of the sigmoid function commonly used in connectionist modeling. Unfortunately, exploring this intriguing possibility would take us far beyond the scope of this paper.

Above I showed that connectionism and AML performed equally well in the toy problem. Why not set aside nonlinearity and instead try to incorporate AML's analysis of homogeneity into the OT model? Given the connectionist-like nature of the OT model, this is, unsurprisingly, rather difficult to conceptualize. AML measures homogeneity by means of overlapping sets and subsets of forms, a device that has no obvious parallel in connectionism or the OT model. This makes it difficult to work out a detailed strategy for making OT work like AML without doing undo violence to its inner OT-nature (and thus possibly alienating the generative linguists whom I hope to count among my audience).

However, a compromise can easily be reached between OT and AML, albeit at a rather superficial level. The final step of the standard AML algorithm, after determining the homogeneous supracontexts and sets of data point/outcome pairs with their associated pointers, is the random selection rule of usage. Since some pointers point to one outcome, others to another, and so on, the rule of usage predicts relative probabilities that are directly proportional to the number of pointers. If one were to write a formula for this, it would appear as follows.

(29) $P(X_1) = |p_{X1}| / [\ |p_{X1}| + |p_{X2}| \ldots + |p_{Xa}|\ ]$,
where $|p_{Xi}|$ represents the number of pointers pointing to outcome X_i.

This is of course identical to the formula given earlier in (15) for the probabilities predicted by the OT model (assuming atomic outcomes). This means that if the AML algorithm is used to generate the proper number of analogical conjoined constraints, OT can be used to generate the probabilities. For example, consider the toy example given in Skousen (1989:22-37), which predicts the probabilities of the outcomes *e* vs. *r* for the context 312, given a set of five data points built of three four-valued features. For each pointer in each homogeneous supracontext generated by the AML algorithm (Skousen 1989:36), posit an OT-like conjoined constraint as given below (W_t represents the target word, here 312).

(30) POINT-OO(W_i, W_j) ∧ IDENT-OO($o(W_j)$,$o(W_t)$)

The first component constraint requires that W_i point to W_j (e.g. 310 points to itself, or 032 points to 212) and the second component requires that the outcomes for the analogical trigger (i.e. the item that the pointer points to) and the target 312 must be completely identical. The first component thus serves as a counting mechanism; the total number of the constraints forcing identity between $o(W_j)$ and o(312) will simply be the number of pointers pointing to W_j, just as in the AML algorithm. That is, if the first component constraint is violated (i.e. there is no such pointer), the conjoined constraint will be violated by every candidate output and so can be ignored (in accordance with Noncommittal Constraint Irrelevance).

The following tableau shows these constraints in action. To save space, I've left out all noncommittal constraints (i.e. those where there is no pointer). A simple application of Anttila's Theorem results in the predicted probabilities P(312*e*) = 4/13 and P(312*r*) = 9/13, precisely as in standard AML (unsurprisingly).

(31)

	o(312) = *o*(310)	*o*(312) = *o*(310)	*o*(312) = *o*(310)	*o*(312) = *o*(310)	*o*(312) = *o*(311)	*o*(312) = *o*(311)	*o*(312) = *o*(311)	*o*(312) = *o*(311)	*o*(312) = *o*(212)	*o*(312) = *o*(212)	*o*(312) = *o*(212)	*o*(312) = *o*(032)	*o*(312) = *o*(032)
312 →*e*					*	*	*	*	*	*	*	*	*
312 →*r*	*	*	*	*									

To make this analysis more palatable to a generativist comfortable with OT, we would need to unpack the constraint POINT. There's no avoiding a full-fledged AML analysis eventually (nor should we necessarily want to, of course), but we might be able to put it off somewhat if we let AML provide us just with the homogeneous supracontexts, and within each, the number of outcomes of each type. Once we know the size *s* of each supracontext and the number *n* of outcomes in it of some type, the number of pointers for this outcome in this supracontext is just *sn*. The total number of pointers pointing to some outcome (i.e. the activation of the outcome node in the OT linear associator) is thus $\Sigma_i\, s_i n_i$. This is the sort of simple mathematics that the OT model could perhaps accommodate, but again exploring this in detail would take us too far afield.

For readers more familiar with AML than with OT, the last part of the above discussion may seem like a trivial parlor trick, but I think there are serious reasons for considering it. First, it represents a perhaps surprising point of contact between two historically different approaches to

language, namely those provided by generative linguistics and by AML. The first step towards cooperation is communication, and I hope that by recognizing this point of contact, scholars of different stripes can learn to use a shared formal language to exchange insights and data. Second, in the previous section I pointed out that it seems difficult to imagine how AML could be modified so that it allows outcomes to be represented with features. Translating the last steps of AML into OT notation serves as a useful aid to the imagination. In fact, all one has to do is modify the constraint in (30) to that in (32), where the second component constraint now refers to a specific feature.

(32) POINT-OO(W_i, W_j) $\wedge$ IDENT-OO($o(W_j)$,$o(W_t)$;F_k)

Making this change would require replacing the standard random rule of usage with the following probability rule (based on the formula in (23)).

(33) $P(o(W_t) = [\alpha_1F_1,..., \alpha_mF_m]) = \Pi_i |p[\alpha_iF_i]| / [|p[\alpha_iF_i]| + |p[\neg\alpha_iF_i]|]$,

where $|p[\alpha_iF_i]|$ represents the number of pointers pointing to an outcome containing feature value $[\alpha_iF_i]$, and $|p[\neg\alpha_iF_i]|$ represents the number of pointers pointing to an outcome containing some other feature value for $[F_i]$.

Again, the consequences of this suggestion are not entirely clear at this point and would require much more thinking than I have space here to work through. I hope, however, that this suggestion sparks some productive thoughts in the reader's mind as well.

6. Beyond analogy

At the end of the preceding section I pointed out one possible way in which the OT model may inspire researchers working on AML. In this section I describe another, namely the ability of the OT model to accommodate certain kinds of nonanalogical factors that conceivably do play a role in human language.

After all, the OT model of analogy described so far only uses one of the two basic types of OT constraints, namely Faithfulness constraints (which I suppose could also include POINT). What about Structure constraints, which require forms to meet universal standards? To a generative linguist, it seems rather foolhardy to claim that all of language can be handled by analogy alone. There are a number of reasons for this, the simplest being that analogy can only work to breathe psychological life into a pattern if there is already something of a pattern there to start with. But where do linguistic patterns come from in the first place? The traditional answer in generative linguistics has been that they come from what OT now calls Structure constraints.

There is another answer, of course: history and physics (or more generally, any set of systematic forces working beyond the confines of a single human brain). Over the past few decades, there has been growing acknowledgement of this alternative answer in generative circles, and some work in OT has used Structure constraints that are explicitly physical in nature (e.g. Flemming 1995, Hayes 1999a, Jun 1995, Silverman 1996, Kirchner 1997, Myers 1997). There's something vaguely disturbing about this, though: Structure constraints are supposed to be (innate) psychological things, so why should they mirror physical forces so exactly? Moreover,

the generativists have never really managed to come up with a convincing reply to critics who suggest that aspects of language (e.g. word-level phonology) are systematic simply because people memorize things that have been molded by extra-mentalist forces over generations of speakers. For example, while it may be true that the [k]~[s] alternation in *electric-electricity* is phonetically natural in some sense, surely this naturalness plays absolutely no role in the minds of modern-day speakers, whose minds are instead occupied with maintaining this pattern through analogy (to the extent that this pattern is psychologically active at all, of course). This modular approach to phonology, where separate subtheories handle the ontogenesis (e.g. physics) and spread (e.g. analogy) of phonological patterns, is completely compatible with the AML program, I think.

Nevertheless, there do seem to be cases where linguistic patterns arise within the minds of speakers, and possibly within the same environment as the mental operations that process analogy. As a case in point, consider patterns that appear to be motivated by innate restrictions on the access, retrieval, and storage of phonological forms. An example of such a pattern is dissimilation. While phonological assimilation can be understood as the semi-fossilization of coarticulation, and hence not fundamentally a psychological phenomenon, dissimilation does not arise through the operation of purely physical forces. Instead, as Ohala (1986) has argued, it requires that listeners in some sense mentally undo perceived coarticulations; when they overshoot, the result is a dissimilation. The generative linguist Kiparsky (1986) endorses this analysis of dissimilation, since it helps explain why dissimilation rules are always lexicalized to some extent and never completely automatic. The natural phonologists Donegan and Stampe (1979) also treat dissimilation as less than purely physical, counting it among fortitions (as opposed to lenitions), which have a perceptual (i.e. psychological) rather than articulatory (i.e. physical) teleology. Taken together, these disparate observers all seem to agree that dissimilations arise not in the outside physical world, but in the mental lexicon.

But this is precisely where analogy occurs as well. No theory of analogy can work without a set of memorized exemplars to analogize from, and Skousen (1989, 1992) even makes memory (and its imperfections) an explicit part of the overall AML approach. In OT terms, this means that there is no theoretical problem with mixing analogical Faithfulness constraints together with Structure constraints, as long as these Structure constraints are motivated by lexical processing rather than physics.

To see how the OT model would do this, I would like to examine a dissimilation pattern first analyzed by Phillips (1981, 1984, 1994) (see also Myers 2000a for a briefer discussion of the same pattern). What makes this pattern particularly interesting is that Phillips (1984) uses it to call attention to an empirical corollary of the above discussion: phonological patterns that are lexically motivated tend to target lower-frequency forms first. The more frequently a form is accessed from memory, the more efficiently it is accessed, and an efficient memory is an accurate one. Hence we do not expect lexical factors to target higher frequency forms. Lower frequency forms, being harder to access, are more subject to whatever plausible hallucinations the memory mechanisms may use to fill in the forgotten holes. This is why analogies tend to affect lower-frequency forms more readily than higher-frequency forms, as I discussed earlier.

With this as background, now consider the pattern. Phillips (1981, 1984, 1994) describes a variable rule in Georgian English whereby the historically older /y/ is optionally deleted after alveolars (including /n/, /d/, and /t/). Crucial for the discussion here is that the probability of this occurring is inversely (not directly) correlated with the frequency of the word. The following table lists an example from each of the five frequency classes that Phillips considers, along with the mean token frequency of each class and the mean probability of y-deletion.

(34)

Example	Mean frequency (tokens)	Probability of y-deletion
new	997.290	0.430
knew	358.380	0.545
numeral	30.290	0.601
neutral	3.594	0.718
nude	0.438	0.744

As Phillips (1984) points out, this is precisely the opposite of the pattern found with phonetically-motivated phonology, which shows positive frequency effects in rate of application or rate of lexical diffusion. For example, the optional dropping of /t/ in words like *mist* during fluent speech, which apparently has an articulatory origin (see e.g. Browman and Goldstein 1990), occurs more often in higher-frequency words than in lower-frequency words (Myers and Guy 1997, Bybee 2000). Other examples of phonetically-motivated phenomena that occur more readily in higher-frequency forms are described in Fidelholtz (1975), Phillips (1984), Kaisse (1985), Hammond (1997), and Bybee (2000), among many other places. As Phillips (1984) and Bybee (2000) have argued, such positive frequency effects are best understood as resulting mainly from physics, not lexical processing (metaphorically speaking, passing words back and forth through the air tends to wear them out). What an analogical model should be able to do, then, is collaborate with lexical factors to create the negative frequency effect seen in Georgian English, but be incapable of interacting directly with the physical forces giving rise to positive frequency effects (such patterns should instead be ascribed to a separate module in the more general theory of phonology).

The OT model meets these criteria. As described in an earlier section, it is capable of describing the fact that lower frequency words make better analogical targets. This ability can also be used to describe the fact that lower frequency words in Georgian English are more likely to show y-deletion. Since y-deletion involves dissimilation, which I argued above is lexically motivated, we may in good conscience posit a Structure constraint to handle it (below I address the question of whether this is in fact necessary rather than merely permissible). For the sake of simplicity, we can use the following constraint (which assumes that alveolars and /y/ are both coronal, i.e. their articulation crucially involves the tongue blade).

(35) *COR-COR: Two adjacent coronals are disallowed (e.g. *[ny]).

Being lexically motivated, this sort of Structure constraint may freely interact with analogical Faithfulness constraints, since both originate in the processes of lexical storage and retrieval. This allows for tableaux like the following, which include both this Structure constraint and a set of Faithfulness constraints that are parochial by tokens. Since the phenomenon is variable within a single dialect, I follow the OT literature on variable phonology (alluded to earlier) and assume that the constraints are freely ranked.

(36) a.

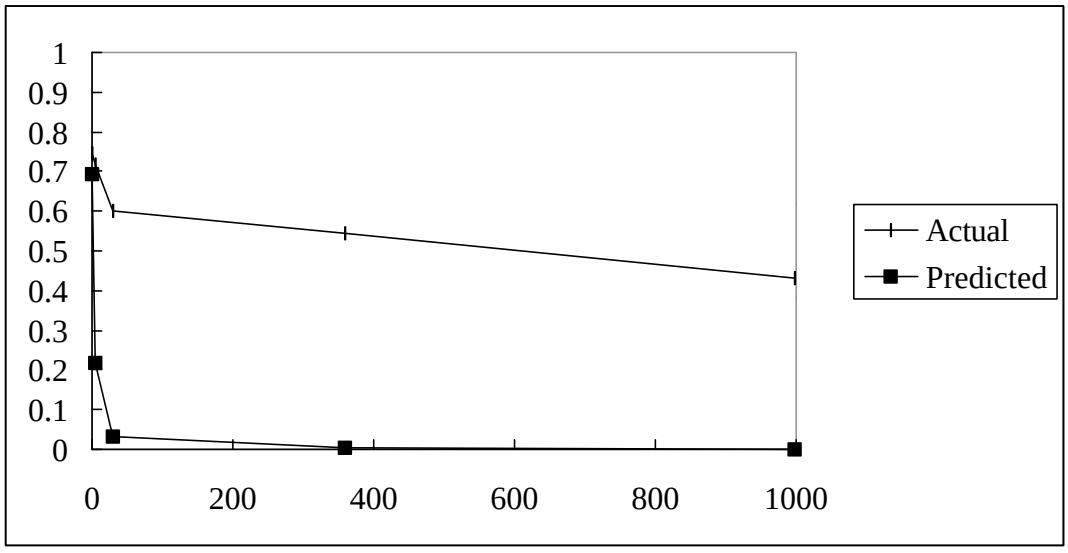
		(1000 of these)		
[nyu]	*COR-COR	IO-*new*	IO-*new*	...
[nyu]	*			...
[nu]		*	*	...

b.

		(1 of these)
[nyud]	*COR-COR	IO-*nude*
[nyud]	*	
[nud]		*

Applying Anttila's Theorem, we derive the probabilities P([nu]) = 0.001 and P([nud]) = 0.500. While as usual the OT model doesn't provide us with particularly accurate numbers, it does capture the essential observation: lower frequency means a higher rate of y-deletion. Both of these points are driven home in the following graph, which shows how far off the OT model is in the specifics while nevertheless resulting in a curve that curves in the correct direction (here I've adjusted the token frequencies to get the best possible values for the OT model at the low end of the frequency distribution).

(37)



In spite of its quantitative inaccuracies, this general approach to structural factors in phonology is satisfying in an important respect: while it can in principle handle negative frequency effects like that just described, it is entirely incapable of handling positive frequency effects (e.g. with t-deletion). If a phonetically motivated Structure constraint were put into an analysis like that sketched in (36), we would falsely predict the frequency effect to be negative as well. This principled weakness is just what we want in an analogical model, which by its very nature is not phonetically motivated. Instead, as noted above, such cases require a separate module of the theory, one independent of analogy.

Nevertheless, I have not made the case that a pattern like y-deletion is necessarily due to a Structure constraint. Surely it would be more parsimonious if an analogical model were always incapable of referring to phonetic entities (like the tongue blade referred to by *COR-COR), regardless of the motivation of the constraints involved. In particular, is it really so inconceivable that y-deletion could itself be due to some lexical process, perhaps even analogy, rather than to a specific constraint of the grammar? Clearly it is conceivable, since Dilworth Parkinson (personal communication) has suggested just such a thing, showing me how AML could derive y-deletion by analogy with higher-frequency words that lack a historical /y/ (e.g. *noon*). This analysis then predicts the negative frequency effects we want, just as with analogy generally. A version of this approach is even possible in the OT model, merely by positing a set of constraints that require *new* and *nude* to share the y-lessness feature with words like *noon*. In any case, frequency effects other than the usual positive one are likely have more than one simple cause. For example, after looking at the same Georgian dialect data, Bybee (2000) came up with a rather different explanation, namely that y-deletion is actually due to borrowing or accommodation to the standard dialect; speakers treat less familiar words less conservatively, allowing them to be replaced with the invading pronunciations. Moreover, in a study of an on-going lexical diffusion in Montreal French, Yaeger-Dror and Kemp (1992) have even discovered a case where frequency appears to play no role at all. Instead the diffusion is affected by semantics (of a curious sort): words keep the older pronunciation if they refer to the "good old days."

Regardless of the final verdict on such cases, I want to leave the reader with a more general lesson: the OT model may represent a case study in how to build a formal model in which analogy can directly interact with (certain) nonanalogical factors (i.e. lexically motivated Structure constraints). Whether or not this is ultimately desirable is an empirical issue, but in the meantime it does seem useful for two reasons. First, generative phonologists have always preferred theories that put the extragrammatical motivations explicitly into the grammar. I think one strategy to help generativists move beyond such theories (which I feel are misguided) may be to get them to examine a model in which extragrammatical motivations (i.e. Structure constraints) are not forbidden a priori, but which predicts that they will behave in very narrowly prescribed ways. Second, research in AML has tended to dismiss too quickly one of the generativist's primary criticisms of analogy, which is that it cannot explain how systematic linguistic patterns arise in the first place. Cases like y-deletion should be collected and carefully examined to determine whether they can all be reanalyzed from a purely analogical perspective, and if not, whether at least some nonanalogical principles of grammar do exist. If such principles are found, something like the OT model described in this paper may help in accommodating them within a mostly analogical formalism.

7. Conclusions

Things are occurring in the Optimality Theory research community that should be of great interest to all those studying analogy. No longer is analogy forbidden in generative linguistics, since there are now widely accepted formal devices that are capable of capturing its essential nature (i.e. exemplar-driven constraints enforcing similarity). My own model may or may not represent a step in the development of a quantitatively successful hybrid between generative and nongenerative approaches to analogy, but it still seems to me to be a rather remarkable fact that something equivalent to a connectionist network can actually be built out of nothing but notions already current in the OT literature. At the very least, I hope that my model inspires generative

linguists to learn more about other explicit models of analogy (especially AML, which deserves far more attention among linguists than it has received). At the same time, I think that researchers in AML and other nongenerative models have something to learn from OT formalism as well, in particular its use of features and its ability to integrate analogy with nonanalogical factors. Analogy is one of the central facts of human language, but it's unlikely to be fully understood without the collaboration of many scholars with different backgrounds and areas of expertise. Perhaps we're now witnessing the beginnings of this collaboration.

References.

Albright, A. To appear. The lexical bases of morphological well-formedness. In S. Benjaballah, W. Dressler, O. E. Pfeiffer, M. D. Voeikova (eds.) *Morphology 2000*. John Benjamins.

Alderete, J. 1999. *Morphologically governed accent in Optimality Theory*. U. Mass. PhD thesis.

Anderson, J. A. 1995. *An introduction to neural networks*. MIT Press.

Anttila, A. 1997. Deriving variation from grammar. In F. Hinskens, R. Van Hout, and W. L. Wetzels [eds] *Variation, change and phonological theory*, 35-68. John Benjamins.

Anttila, A., and Cho, Y-M. 1998. Variation and change in Optimality Theory. *Lingua* 104:31-56.

Baayen, R. H. 1995. Review of *Analogy and structure*. *Language* 71:390-396.

Balari, S., Marí n, R., and Vallverdú, T. 2000. Implicational constraints, defaults and markedness. Universitat Autònoma de Barcelona ms. ROA.

Benua, L. 1995. Identity effects in morphological truncation. In J. Beckman, L. Walsh Dickey and S. Urbanczyk [eds] *University of Massachusetts Occasional Papers in Linguistics* 18:77-136.

Benua, L. 1997a. Affix classes are defined by Faithfulness. *University of Maryland Working Papers in Linguistics* 5:1-26.

Benua, L. 1997b. *Transderivational Identity: phonological relations between words*. U.Mass. PhD diss.

Boersma, P. 1998. *Functional Phonology: Formalizing the interactions between articulatory and perceptual drives.* U. Amsterdam doctoral dissertation. The Hague: Holland Academic Graphics.

Boermsa, P. and Hayes, B. 2001. Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32:45-86.

Booij, G. 1997. Non-derivational phonology meets lexical phonology. In I. Roca (ed.) *Derivations and constraints in phonology*, 261-288. Clarendon Press.

Browman, C. P., and Goldstein, L. (1990) Tiers in articulatory phonology, with some implications for casual speech. In J. Kingston and M. E. Beckman (eds.) *Papers in laboratory phonology 1: between the grammar and physics of speech.* Cambridge University Press, 341-376.

Burzio, L. 1997a. Strength in numbers. In V. Miglio and B Morén (eds.) *University of Maryland Working Papers in Linguistics* 5:27-52.

Burzio, L. 1997b. Surface constraints versus underlying representation. In Durand and Laks, 97-122.

Burzio, L. 1999. Surface-to-surface morphology: when your representations turn into constraints. Johns Hopkins University Department of Cognitive Science ms. Presented at

the 1999 Maryland Mayfest, University of Maryland, College Park.
Burzio, L. 2000. Cycles, non-derived environment blocking, and correspondence. In J. Dekkers and F. van der Leeuw (eds.) *Optimality Theory: syntax, phonology and acquisition*. Oxford University Press.
Burzio, L. To appear. Missing players: phonology and the past-tense debate. *Lingua*.
Bybee, J. L. 2000. The phonology of the lexicon: evidence from lexical diffusion. In M. Barlow and S. Kemmer (eds.) *Usage-based models of language*, 65-85. Stanford: CSLI.
Crowhurst, M. and Hewitt, M. 1997. Boolean operations and constraint interactions in Optimality Theory. University of North Carolina at Chapel Hill and Brandeis University ms. ROA.
Derwing, B. L., and Skousen, R. 1994. Productivity and the English past tense: testing Skousen's analogy model. In S. D. Lima, R. L. Corrigan, and G. K. Iverson [eds.] *The Reality of Linguistic Rules*, 193-218. John Benjamins.
Donegan, P. J. and Stampe, D. 1979. The study of natural phonology. In D. A. Dinnsen [ed.] *Current Approaches to Phonological Theory*, 126-173. Indiana University Press.
Durand, J. and Laks, B. [eds.] 1997. *Current trends in phonology: models and methods*. Salford: University of Salford.
Fidelholtz, J. L. 1975. Word frequency and vowel reduction in English, *Chicago Linguistics Society* **11**.200-213.
Flemming, E. 1995. *Perceptual features in phonology*. UCLA PhD thesis.
Golston, C. 1996. Direct Optimality Theory: Representation as pure markedness." *Language* 72:713-748.
Green, A. D. 2001. The tense-lax distinction in English vowels and the role of parochial and analogical constraints. *Linguistics in Potsdam* 15. ROA.
Hale, M., Kissock, M. and Reiss, C. 1998. Output-output correspondence in Optimality Theory. Proceedings of WCCFL XVI. Stanford: CSLI Publications.
Hammond, M. 1995. There is no lexicon! University of Arizona ms. ROA.
Hammond, M. 1997. Lexical frequency and rhythm. University of Arizona ms. ROA.
Hayes, B. 1999a. Phonetically-driven phonology: the role of Optimality Theory and inductive grounding. In M. Darnell, E. Moravscik, M. Noonan, F. Newmeyer, and K. Wheatly (eds.) *Functionalism and Formalism in Linguistics, Vol. 1: General Papers*, 243-285. John Benjamins.
Hayes, B. 1999b. On the richness of paradigms, and the insufficiency of underlying representations in accounting for them. UCLA ms.
Hayes, B., and MacEachern, M. 1998. Folk verse form in English. *Language* 74:473-507.
Jun, J. 1995. *Perceptual and articulatory factors in place assimilation: an Optimality Theoretic approach.* ULCA PhD thesis.
Kaisse, E. M. 1985. *Connected Speech: The Interaction of Syntax and Phonology*. Academic Press.
Kenstowicz, M. 1995. Cyclic vs. non-cyclic constraint evaluation. *Phonology* 12:397-436.
Kenstowicz, M. 1997. Base-Identity and Uniform Exponence: alternatives to cyclicity." In Durand and Laks, 363-394.
Kiparsky, P. 1978. Analogical change as a problem for linguistic theory. Reprinted in P. Kiparsky [ed] 1982. *Explanation in Phonology*. Foris Publications: Dordrecht.
Kiparsky, P. 1986. Commentary on Ohala 1986. In Perkell and Klatt.
Kiparsky, P. 1988. Phonological change. In F. Newmeyer [ed.] *Cambridge Survey of Linguistics, vol. I* , 363-410. Cambridge University Press: Cambridge.
Kiparsky, P. 1993. Variable rules. Handout for Rutgers Optimality Workshop 1.

Kirchner, R. 1997. Contrastiveness and faithfulness. *Phonology* 14:83-111.
Kirchner, R. 1999. Preliminary thoughts on 'phonologization' within an exemplar-based speech processing system. University of Alberta ms. ROA.
McCarthy, J. and Prince, A. 1993a. *Prosodic Morphology I: Constraint interaction and satisfaction*. Forthcoming from MIT Press.
McCarthy, J. and Prince, A. 1993b. Generalized alignment. In G. Booij and J. van Marle [eds.] *Yearbook of Morphology*. Kluwer. 79-153.
McCarthy, J. and Prince, A. 1995. Faithfulness and reduplicative identity. In J. Beckman, L. Walsh Dickey and S. Urbanczyk [eds.] *University of Massachusetts Occasional Papers in Linguistics* 18:249-384.
Myers, J. 1997. Canadian Raising and the representation of gradient timing relations. *Studies in the Linguistic Sciences* 27:169-184.
Myers, J. and Guy, G. R. 1997. Frequency effects in Variable Lexical Phonology. *University of Pennsylvania Working Papers in Linguistics* **4**.215- 228.
Myers, J. 2000a. Analogy and optimality. National Chung Cheng University ms.
Myers, J. 2000b. Variable constraint ranking in Optimality Theory. National Chung Cheng University ms.
Nagy, N. and Reynolds, B. 1997. Optimality Theory and word-final deletion in Faetar. *Language Variation and Change* 9:37-55.
Ohala, J. J. 1986. Phonological evidence for top-down processing in speech production. In Perkell and Klatt. 386-401.
Perkell, J. S. and Klatt, D. H. [eds.] 1986. *Invariance and variability in speech processes*. Lawrence Erlbaum Associates: Hillsdale, N.J.
Phillips, B. 1981. Lexical diffusion and Southern tune, duke, news. *American Speech* 56:72-78.
Phillips, B. 1984. Word frequency and the actuation of sound change. *Language* 45:9-25.
Phillips, B. 1994. Southern English glide deletion revisited. *American Speech* 69:115-127.
Plunkett, K. and Elman, J. L. 1997. *Exercises in rethinking innateness: a handbook for connectionist simulations.* Cambridge, MA: MIT Press.
Prince, A. and Smolensky, P. 1993. *Optimality Theory: Constraint interaction in generative grammar*. Rutgers University and University of Colorado ms.
Prince, A. and Smolensky, P. 1997. Optimality: from neural networks to universal grammar. *Science* 275:1604-1610.
ROA. Rutgers Optimality Archive: http://roa.rutgers.edu/
Russell, K. 1995. Morphemes and candidates. U. of Manitoba ms. ROA.
Russell, K. 1999. MOT: Sketch of an OT approach to morphology. U. of Manitoba ms. ROA.
Samek-Lodovici, V. and Prince, A. 1999. Optima. University College, London, and Rutgers University, New Brunswick, ms. ROA.
Silverman, D. 1996. Voiceless nasals in auditory phonology. *Proceedings of the Berkeley Linguistic Society* 22:364-374.
Skousen, R. 1989. *Analogical modeling of language.* Kluwer Academic Publishers.
Skousen, R. 1992. *Analogy and structure*. Kluwer Academic Publishers.
Smolensky, P. 1995. On the internal structure of the constraint component *Con* of UG. Paper presented at UCLA.
Steriade, D. 2000. Paradigm uniformity and the phonetics-phonology boundary. In M. B. Broe and J. B. Pierrehumbert (eds.) *Papers in Laboratory Phonology V: Acquisition and the Lexicon*, 313-334. Cambridge University Press.
Steriade, D. 1999a. Lexical conservativism in French adjectival liaison. In B. Bullock, M.

Authier, L. Reed (eds.) *Formal Perspectives in Romance Linguistics*. John Benjamins.
Steriade, D. 1999b. Lexical conservativism and the notion *base of affixation*. UCLA ms.
Tversky, A. (1977) Features of similarity. *Psychological Review* 8:327-352.
Yaeger-Dror, M., and Kemp, W. 1992. Lexical classes in Montreal French: the case of (E:). *Language and Speech* 35:251-293.