Output-Driven Learning, Restrictiveness and an Inductive Leap: learning stressless prosodic systems from distributional evidence

Morgan C. Moyer

May 15, 2017

1 The Challenge of Restrictiveness: learning the most restrictive grammar in the face of a subset-superset relation

A major and notorious challenge to language learning arises when the learner encounters a set of data that is consistent with two different grammars, and the language generated by one grammar forms a subset of the language generated by the other. If the learner is learning the subset, no observed word (i.e. postitive evidence) will be inconsistent with the superset. This phenomenon is known as *the Subset Problem* (Angluin 1980, Baker 1979). However, the superset grammar generates words that the subset grammar does not. Such evidence indicating those words are not in the language would distinguish the superset language from the subset on the basis of consistency. However, children rarely if ever receive robust negative evidence; learning occurs largely on the basis of positive evidence (Brown and Hanlon 1970).

The traditional subset problem defined a relation between the phonotactic forms of two languages. One language forms a *phonotactic subset* of another if the second contains all the forms of the first, plus additional ones. As Tesar (2014) shows, one language can also be a *paradigmatic subset* of another. One language can be a paradigmatic subset of another when the set of possible morpheme behaviors of one language forms a subset of the possible morpheme behaviors of the second.

Berwick (1982) proposed a simple approach to the problem: when forced to choose between two grammars, one of which is a subset of the other, choose the subset grammar to avoid a grammar that overgenerates. As noted in Prince and Tesar (2004), this simple approach is rather complicated to implement computationally: the learner would have to pairwise compare the restrictiveness of all the possible grammars. Unfortunately, the number of possible grammars is most likely too large to conduct pairwise comparisons.

However, a coherent idea underlies Berwick's approach. Preferring the subset grammar means avoiding overgeneration of phenomena that the grammar allows. The learner should be somewhat economical in selecting a hypothesis consistent with the data: it should be as restrictive as possible while maintaining data consistency. Choosing the most restrictive grammar means choosing the grammar that generates all and only the grammatical forms of the language.

Previous approaches within the framework of Optimality Theory observed the connection between restrictiveness and a ranking bias where all markedness constraints dominate all faithfulness constraints (Bernhardt and Stemberger (1998), Demuth (1995), Gnanadesikan (2004), Levelt (1995), Sherer (1994), Smolensky (1996), Smolensky (1996), Oostendorp (1995)). Markedness constraints lead to more restrictive grammars because they assign violations to marked ouputs, forcing neutralization, and thereby restrict ouput inventories. Faithfulness constraints have the opposite effect: they expand output inventories by preserving underlying distinctions.

Prince and Tesar (2004) propose a concrete implementation of this restrictiveness bias in OT using an algorithm which operates over a set of winner-loser pairs to generate a constraint ranking.

This is called Biased Constraint Demotion (BCD). The winner-loser pairs encode ranking information that learner collects over the course of learning. The collection represents the learner's current hypothesis for the grammar, and can flexibly change as new data is observed by the addition of new winner-loser pairs. At any point in learning, a ranking may be constructed using the bias that ranks markedness constraints before faithfulness constraints.

Tesar (2014) proposed a learning algorithm that exploits logical structure in the space of possible inputs (output-drivenness) to incrementally learn underlying forms and the constraint rankings which determine those forms, while maintaining as restrictive a hypothesis as possible. The algorithm enforces restrictiveness by using the methods of Prince and Tesar (2004), in addition to a method that attempts to preserve input neutralization through setting underlying feature values only as necessary to account for surface contrasts. This is the Output-Driven Learner (ODL). Tesar (2014) defined a typology (SL) of 24 languages modeling the interaction of stress and vowel length. When the ODL was applied to this typology, all 24 languages were successfully learned despite two cases of paradigmatic subsets.

The current paper explores a new output-driven system (NSL) which gives rise to significant challenges to restrictiveness. The system modifies Tesar's to model stress systems that allow stressless outputs. Section 2 presents both systems, and includes the empirical and computational motivation for the implementations in the new system. First, the properties of paradigmatic subsets as identified first by Tesar (2014) are presented in Section 2.1.3. In Section 2.4.3 and Section 2.5.3, we introduce two cases of paradigmatic subsets which have different properties than those described in Section 2.1.3. Section 3 pertains to output-drivenness, the algorithm, and Section 3.3 how it enforces restrictiveness. Section 4 discusses the successful learning of the language from SL presented in Section 2.1.1.

However, the ODL as formalized in Tesar (2014) fails to learn the more restrictive grammar when faced with the two cases of paradigmatic subsets presented in Sections 2.4.3 and 2.5.3. Sections 5 and 6 present the failed learning of these two NSL languages.

Sections 5.2 and 5.3 present a successful solution to restrictiveness for the first case. This solution attempts to exploit the phonotactic distribution of observed words in the language. If the learner can make an inductive leap and assume that the words observed thus far are all and only the phonotactic forms generated by the grammar, it then follows that all unobserved but possible output forms must be neutralized by the grammar to the observed words. If only one mapping from a given unobserved word to each observed word is consistent with the set of winner-loser pairs that constitute the learner's evidence for a grammar, then the learner can find explicit evidence of non-phonotactic ranking information that enforces such a neutralization, information that is unavailable on the basis of the observed words alone. The success of this solution lends support and validity to the pursuit of absent data for language learning. The paper ends with a discussion of the results and implications of the proposed solution.

2 The Linguistic Systems

This section presents the two linguistic systems which will be discussed in this paper. The baseline system is Tesar's (2014) Stress/Length (SL), presented first below. That system was modified with the aim to model stress language systems allowing stressless words. This section presents both systems in order to discuss and provide concrete examples of paradigmatic subsets, the properties of such cases, and how they pose a challenge to restrictiveness.

2.1 The Stress/Length System (SL)

Words are bisyllabic, consisting of an onset [p] in the first syllable (the root), and [k] in the second syllable (the suffix). Each syllable contains of a single vowel. The vowels are specified for

two features, length (+/-) and stress (+/-). In keeping with Richness of the Base, there are eight morphemes total, for each possible combination of underlying features.

(1) Morphemes in the lexicon of SL
   a. Root: r1 = /pa/, r2 = /pa:/, r3 = /pá/, r4 = /pá:/
   b. Suffix: s1 = /ka/, s2 = /ka:/, s3 = /ká/, s4 = /ká:/

For each morpheme, there are four possible underlying forms. For two syllable words, there are 16 viable input forms for every combination of morphemes in the correct order, and 65,536 possible lexica from those 8 morphemes (4^8). In the SL system, GEN restricts outputs forms to only those that are culminative (i.e., have exactly one main stress), thus the following violating forms are excluded: [paka], [pa:ka], [paka:], [pa:ka:], and [páká], [pá:ká], [páká:], [pá:ká:]. There are six constraints, listed and defined below.

(2) The Constraints of SL
   a. WSP: long vowels must be stressed (weight-to-stress principle)
   b. MAIN LEFT: main stress on the initial syllable
   c. MAIN RIGHT: main stress on the final syllable
   d. NO LONG: no long vowels
   e. ID[STRESS]: IO correspondents have equal stress value
   f. ID[LENGTH]: IO correspondents have equal length value

These constraints define 720 possible total rankings (6!), but a typology of only twenty-four languages.

Of those 24 languages, 22 were successfully learned using only the restrictiveness enforcing Biased Constraint Demotion of Prince and Tesar (2004). The other two languages, L8 and L17, required an additional mechanism called Fewest Set Features to successfully complete learning. These failed because they formed a paradigmatic subset to another successfully learned language. This additional mechanism will be discussed in more detail later in the paper, but the next sections illustrates the properties of the subset L8 and its superset L7.

2.1.1 Language L8, the subset language

A ranking generating L8 is given in (3a), and the language itself in (3b):

(3) a. WSP >> ID[Length]>> NoLong >> MR >> ML >> ID[Stress]
   b. Language L8

	r1 = /pa/	r2 = /pa:/	r3 = /pá/	r4 = /pá:/
s1 = /ka/	paká	paká	paká	pá:ka
s2 = /ka:/	paká:	paká:	paká:	paká:
s3 = /ká/	pá:ka	pá:ka	paká	pá:ka
s4 = /ká:/	paká:	paká:	paká:	paká:

L8 contains significant morphemic neutralization: across roots r1/r3, r2/r4, and across suffixes s1/s3, and s2/s4. This is because of the low position of ID[Stress], and the higher ranking of WSP and ID[Length]: stress will be attracted to long syllables, which will surface faithful. If a word has two morphemes underlyingly short, stress surfaces by default on the left. Unstressed vowels will surface short.

We can show the morphemic neutralizations in the compressed chart below:

(4) L8 compressed

	/pa/, /pá/	/pa:/, /pá:/
/-ka/, /-ká/	paká	pá:ka
/-ka:/, /-ká:/	paká:	paká:

As represented in the chart, stress is not contrastive for any morpheme. Though the above table has both stressed and unstressed morphemes, it is important to note that either stress value is consistent with the phonotactic forms, and the chart could have been written as two charts, one with all morphemes underlyingly stressed, and the other with them unstressed. This point will be returned to later on.

The phonotactic inventory of L8 has only three words: paká, paká: and pá:ka.

2.1.2 Language L7, the superset language

A ranking generating L7 is given in (5a) and the language itself in (5b):

(5) a. WSP >> ID[Stress] >> ID[Length] >> NoLong >> MR >> ML

b. Language L7

	r1 = /pa/	r2 = /pa:/	r3 = /pá/	r4 = /pá:/
s1 = /ka/	paká	pá:ka	páka	pá:ka
s2 = /ka:/	paká:	paká:	páka	pá:ka
s3 = /ká/	paká	paká	paká	pá:ka
s4 = /ká:/	paká:	paká:	paká:	paká:

The rich input base contains four roots and four suffixes. Unlike L8, L7 has no total neutralization, as every pair of roots and pair of suffixes behaves differently in at least one environment. Note the high ranking of both faithfulness constraints, which contributes to this effect. The phonotactic inventory of L7 has four words: paká, paká:, pá:ka, and páka.

2.1.3 L8 is a paradigmatic subset of L7

L7 has all the phonotactic forms of L8, in addition to [páka]. L8 is thus a phonotactic subset of L7. Words of L8 only have stress when the initial vowel is long. L7 words can have initial stress with both long and short (initial) vowels.

These facts are reflected when we compare rankings that generate the two languages:

(6) L8: WSP >> ID[Length]>> NoLong >> MR >> ML >> **ID[Stress]**

(7) L7: WSP >> **ID[Stress]** >> ID[Length]>> NoLong >> MR >> ML

The difference between lies in the location of ID[Stress]: in L8 it is ranked last, while in L7 it is ranked second from the top. This accounts for the more restricted distribution of stress in L8, but the freer, less restricted distribution in L7.

L8 is also a paradigmatic subset of L7. In fact, Tesar notes there are several ways to "project" L8 into L7. If we represent the underlying morphemes of L8 as unstressed, we can see that the mapping from inputs to outputs in L8 forms a subset of the mapping in L7.

(8) L8 compressed table with morphemes underlyingly stressed

	/pá/	/pá:/
/-ká/	paká	pá:ka
/-ká:/	paká:	paká:

(9) L7 with L8-equivalent shaded subset projection

	r1 = /pa/	r2 = /pa:/	r3 = /pá/	r4 = /pá:/
s1 = /ka/	paká	pá:ka	páka	pá:ka
s2 = /ka:/	paká:	paká:	páka	pá:ka
s3 = /ká/	paká	paká	paká	pá:ka
s4 = /ká:/	paká:	paká:	paká:	paká:

All the morphemes of L8 are contained in the shaded forms of L7: the surface alternations of morphemes in L8 are consistent with those of L7. Thus, any data set consistent with the first language will be consistent with the second. If the neutralizing morphemes of L8 have the same value for stress, then a stress contrast is hidden in the lexicon.

However, if the underlying morphemes of L8 are represented as unstressed underlyingly, a different subset with L7 is formed.

(10) L8 compressed table, with morphemes underlyingly unstressed

	/pa/	/pa:/
/-ka/	paká	pá:ka
/-ka:/	paká:	paká:

(11) L7 with second L8-equivalent subset shaded

	r1 = /pa/	r2 = /pa:/	r3 = /pá/	r4 = /pá:/
s1 = /ka/	paká	pá:ka	páka	pá:ka
s2 = /ka:/	paká:	paká:	páka	pá:ka
s3 = /ká/	paká	paká	paká	pá:ka
s4 = /ká:/	paká:	paká:	paká:	paká:

Again, we see the stressed morphemes of L8 form two different subsets with L7: in (11), as in (9), the projection into L7 retains the same neutralization of stress across roots, while maintaining contrast in length.

However, if the morphemes of L8 have different values for stress underlyingly, we see yet a different subset relation emerge with L7:

(12) L7 with superficially L8-equivalent subset shaded

	r1 = /pa/	r2 = /pa:/	r3 = /pá/	r4 = /pá:/
s1 = /ka/	paká	pá:ka	páka	pá:ka
s2 = /ka:/	paká:	paká:	páka	pá:ka
s3 = /ká/	paká	paká	paká	pá:ka
s4 = /ká:/	paká:	paká:	paká:	paká:

The shaded regions of (12) now do not contrast in length on the root, but in stress: both roots are underlyingly long and never surface faithful to length unless stressed. The suffixes still contrast only in length. L8 does not have stress contrast, but L7 does. The shaded words of (9) and (11) are consistent with stress neutralization in L8, but the shaded words of (12) contrast in stress underlyingly and on the surface in L7.

If we take the ranking generating L8 from (3a), using the underlying forms which map to the shaded words in (12), we see that the resulting paradigm contains a different mapping from those inputs:

(13) Words using the ranking of L8: no length contrast in roots so identical root behavior

	/pa:/	/pá:/
/-ká/	pá:ka	pá:ka
/-ká:/	paká:	paká:

We can conclude that the underlying features of the subparadigm in (12) are in fact inconsistent with the ranking of L8, because the underlying values in that subparadigm map to different surface forms with the ranking of L8. These feature values are not contrastive for L8, but are for L7.

We have just witnessed our first example of a paradigmatic subset, a relation defined by the morphemic behavior between two languages. L8 has two contrasting roots, one which always surfaces short and unstressed, and the other which surfaces either short and unstressed in one environment or long and stressed in another environment. L7 has three contrasting roots, one which always surfaces short and unstressed, another which surfaces either short and stressed or short and unstressed, and the third which surfaces stressed and long. This last root behavior is absent from L8.

The projection of L8 into L7 demonstrates the paradigmatic subset relation between the two languages. Though there were multiple ways of projecting L8 into L7, one projection is in fact inconsistent with the ranking generating L8: the surface root contrast in length of L8 could be accounted for using underlying forms that contrasted in stress, but these underlying forms are consistent only with the grammar of L7, as shown in (13).

However, the learner does not have access to this information. It only encounters surface forms, and must determine on the basis of those forms their underlying values and the ranking which maps underlying to observed forms. From the learner's perspective then, there are two ways to account for the surface alternation in L8. In (9) and (11), the roots of L8 must contrast in length but agree in stress. In (12), the roots contrast in stress, and the +stress morpheme must also be +long. A restrictive feature setting approach would choose the latter solution, because it allows for more input neutralization. This fact will enter into the discussion of the Fewest Set Features mechanism, discussed later in Section 3.3.2.

Before moving on to the next examples of paradigmatic subsets, we first motivate the need for the new system which will generate those examples. This motivation is provided in the next section, and the new system will be introduced following that.

2.2 Motivation for a new system

This section articulates the reasons motivating the proposed modifications to Tesar (2014). Little research has been done on the properties of systems that the Output-Driven Learner can learn– apart from having the property of output-drivenness. Output-drivenness will be introduced in Section 3.

Importantly, the goal is to introduce a new Output-Driven system which allows stressless output words, by relaxing restrictions on the obligatoriness of main stress. Arguably, some languages do allow stressless lexical words. For example, Cayuga, Seneca, Sierra Miwok, Yup'ik, Kinyambo and Indonesian all purportedly permit stressless words (Hayes 1995:25, Athanasopoulou, et al, submitted).

To give a complete phonological analysis, the natural path would be to explicitly represent prosodic structure in the new system, by representing a metrical head as a binary feature in addition to the stress and length features present in SL, in addition to adding constraints which refer to metrical structure.

However, introducing metrical structure adds hidden linguistic structure in the output, and thus creates significant potential for ambiguity. To see the full effect of such a proposal, consider an example of a tri-syllabic word with medial stress from Tesar (2004):

(14) $\sigma\acute{\sigma}\sigma$
  a. $(\sigma\acute{\sigma})\sigma$
  b. $\sigma(\acute{\sigma}\sigma)$

(14a) and (14b) describe two compatible structural analyses of the overt form that the learner hears. (a) presents the first two syllables grouped together to form an iamb, while (b) presents a second grouping to form a trochee. The learner cannot determine which analysis is correct just on the basis of the overt form alone. The grammar makes reference to the metrical structure which cannot be observed by the learner.

Such ambiguity also arises with the computation of secondary stress, which lies in the interactions of multiple feet (the learnability of which is the subject of Akers' 2012 dissertation). Allowing multiple realizations of stress in an output will likewise create a structural ambiguity.

In Optimality Theory, one may refer to ambiguity of the input (lexical ambiguity), or ambiguity of the output (structural ambiguity). The leading approach to learning in the face of structural ambiguity is Tesar (2004), while Akers (2012) deals with the joint learning of lexical and structural ambiguity. In either case, the learning of structurally ambiguous forms involves maintaining and processing multiple representations in parallel—an effective but far from trivial process. We thus maintain only lexical ambiguity (following Tesar 2014) by avoiding multiple realizations of high tones/stresses and direct reference to prosodic structure, in order to side-step such complexity for the purpose of this project.

2.3 The New Stress/Length System (NSL)

In Tesar (2014), the surface realization of stress implicitly indicated the position of the metrical head. We will maintain this idealization rather than explicitly introduce metrical structure. There are two steps to relaxing obligatory stress: GEN is modified to allow outputs without main stress, and a new constraint is added to CON which assigns violations to those stressless outputs. We define our

new constraint below, consistent with an intuitive notion of obligatory stress marking[1]:

(15) a. OBLIG: Outputs must have *exactly one* main stress.
b. CON (n=7): OBLIG, WSP, MAIN LEFT, MAIN RIGHT, NO LONG, ID[STRESS], ID[LENGTH]

The addition of this constraint to the six above yields seven constraints total, which generate 5,040 possible rankings (7!), but only 62 distinct languages.

As noted we maintain a ban on multiple realizations of stress as a restriction in GEN, therefore blocking four outputs: [páká], [pá:ká], [páká:], [pá:ká:]; but allowing four previously blocked outputs: [paka], [pa:ka], [pa:ka], [pa:ka:]. In SL, all eight of these were impossible outputs. In keeping with Richness of the Base, there are eight possible underlying morphemes, for each combination of stress and length feature values:

(16) *Morphemes in the lexicon of NSL*
a. Root: r1 = /pa/, r2 = /pa:/, r3 = /pá/, r4 = /pá:/
b. Suffix: s1 = /ka/, s2 = /ka:/, s3 = /ká/, s4 = /ká:/

Besides modifications to GEN and CON, the space of possible underlying forms in this system is identical to the SL system.

Traditionally in the literature on alignment, the generalized alignment constraints commonly called ALL-FEET-RIGHT and ALL-FEET-LEFT are acknowledged foot antagonists (McCarthy & Prince 1993). These constraints align, for each foot, the right or left foot edge with some prosodic word (McCarthy & Prince 1993). At most one foot can be fully aligned with a right or left edge of a word, so additional feet not aligned to an edge will incur violations from these constraints. These additional violations can be avoided by minimizing or all together avoiding additional feet. Insofar as GEN might allow candidates to vary on the footing of syllables, the two alignment constraints will mostly prefer candidates with fewer feet. They are thus foot antagonists.

In contrast, ML and MR as formalized in SL are not foot antagonists, since GEN does not allow any candidates without main stress (or with two main stresses). All candidates receive a violation from one constraint or the other if main stress is not at the relevant word edge, as there are no stressless candidates for either constraint to prefer over a stressed one. By relaxing the obligatoriness of stress in NSL's GEN, ML and MR become stress antagonists. Though the constraints actually refer to stress rather than directly to foot heads, the point is the same. GEN allows candidates without a main stress, so it is possible not to have a main stress and fully (vacuously) satisfy ML and MR.

Thus, given these facts about NSL's ML and MR, we find two languages in the new typology where there are no phonotactic words with stress: Languages L53 and L62 exhibit the logical consequences of the antagonism discussed here. In these languages, ML and MR are ranked highest together in the constraint ranking with the effect that no candidate with stress will be more optimal than a stressless one.

The next two sections illustrate two cases of paradigmatic subsets in the NSL system. The two cases can be distinguished from each other by the phonotactic behavior between the subset and superset languages: one case constitutes a phonotactic subset relation (discussed in Section 2.4), and the other a phonotactic identity (discussed in Section 2.5).

[1]Hyman (2009) separates out two (non-OT constraint) sides of this property of having exactly one main stress: Obligatoriness, "Every lexical word has at least one syllable marked for the highest degree of prosodic prominence"; and Culminativity, "Every lexical word has at most one syllable marked for the highest degree of prosodic prominence," (Hyman 2009:217). The constraint as formalized in (15a) is functionally akin to an OT constraint version of Hyman's Obligatoriness because it penalizes outputs with no main stress. We word it slightly differently as this system's GEN does not generate outputs with multiple stresses, so there is no need to differentiate between more than one and at least one main stress.

2.4 Case 1: a Phonotactic Subset

2.4.1 Language L32, the subset language

Language L32 has final stress, but allows two stressless forms to surface when the suffix is underlyingly short and unstressed. The description of Language L32 is presented in (17) along with a ranking that generates the language:

(17) a. MR >> ID[Length] >> {NoLong, WSP} >> ID[Stress] >> {ML, Oblig}
b. Language L32

	r1 = /pa/	r2 = /pa:/	r3 = /pá/	r4 = /pá:/
s1 = /ka/	paka	pa:ka	paka	pa:ka
s2 = /ka:/	paká:	pa:ká:	paká:	pa:ká:
s3 = /ká/	paká	pa:ká	paká	pa:ká
s4 = /ká:/	paká:	pa:ká:	paká:	pa:ká:

Several morphemes neutralize in this language: r1/r3, r2/r4, and s2/s4. There is no stress contrast across any root, partly a consequence of MR dominating all other constraints. Neutralization between s2 and s4 is due to the fact that WSP dominates ID[Stress]: stress surfaces on long suffixes regardless of underlying stress value. We can rewrite the morpheme behavior below:

(18) L32 compressed

	/pa/, /pá/	/pa:/, /pá:/
/-ka/	paka	pa:ka
/-ka:/, /-ká:/	paká:	pa:ká:
/-ká/	paká	pa:ká

This chart shows explicitly the morphemic neutralizations in L32.

The phonotactic inventory of L32 has six words:

(19) L32 Phonotactic Inventory: paka, pa:ka, paká:, pa:ká:, paká, pa:ká

2.4.2 Language L58, the superset language

Language L58 has final stress but allows stressless forms to surface when unstressed suffixes differ in length. A ranking generating L58 is shown in (20a), and the language itself is shown in (20b):

(20) a. MR >> ID[Stress] >> {ML, Oblig} >> ID[Length] >> {NoLong, WSP}
b. Language L58

	r1 = /pa/	r2 = /pa:/	r3 = /pá/	r4 = /pá:/
s1 = /ka/	paka	pa:ka	paka	pa:ka
s2 = /ka:/	paka:	pa:ka:	paka:	pa:ka:
s3 = /ká/	paká	pa:ká	paká	pa:ká
s4 = /ká:/	paká:	pa:ká:	paká:	pa:ká:

L58 has neutralization across roots r1/r3 and r2/r4, reflective of MR's dominance of ID[Stress]. L58 has no neutralization across suffixes: each pair of suffixes behaves differently in at least one environment. This is the consequence of both MR and ID[Stress]'s high positions in the ranking: final stress on an output will satisfy faithfulness to stress in the suffix. We can compress the morpheme behaviors of L58 in the following chart:

(21) Language L58 Compressed

	/pa/, /pá/	/pa:/, /pá:/
/ka/	paka	pa:ka
/ka:/	paka:	pa:ka:
/ká/	paká	pa:ká
/ká:/	paká:	pa:ká:

The phonotactic inventory of L58 has eight forms.

(22) L58 Phonotactic Inventory: paka, pa:ka, paka:, pa:ka:, paká:, pa:ká:, paká, pa:ká

2.4.3 L32 is a paradigmatic and phonotactic subset of L58

The phonotactic forms of L32 are a subset of the phonotactic forms of L58. L58 contains all the surface forms of L32 with the addition of two: [paka:] and [pa:ka:], the outputs of r1s2/r3s2 and r2s2/r4s2, respectively. Recall that in both L32 and L58, stress neutralizes in the root, but only L58 shows four distinct suffix behaviors while L32 neutralizes long suffixes. If we compare the morphemic behavior of the two languages, we see that L32 forms a paradigmatic subset of L58:

(23) L32 compressed

	/pa/, /pá/	/pa:/, /pá:/
/-ka/	paka	pa:ka
/-ka:/, /-ká:/	paká:	pa:ká:
/-ká/	paká	pa:ká

(24) L58 with L32-equivalent subset shaded

	/pa/, /pá/	/pa:/, /pá:/
/ka/	paka	pa:ka
/ka:/	paka:	pa:ka:
/ká/	paká	pa:ká
/ká:/	paká:	pa:ká:

The shaded forms in (24) correspond to the forms in (23). In L58 suffixes never surface unfaithfully, but L32 globally neutralizes two suffixes: L32 lacks long unstressed suffixes on the surface. L58 is less restrictive of stressless forms. Unlike in the case of L7 and L8 of the SL system, regardless of the underlying value assigned to the neutralizing long suffixes (whether stressed or unstressed) there is only one "projection" of L32 into L58. Quite possibly this is due to the fact that L58 already contains significant morphemic neutralization, there is no room for multiple projections of L32 into L58.

In L32, s2 neutralizes with s4 because ID[Length] requires s2 to surface long, then WSP requires it to be stressed. These two constraints dominate ID[Stress], which would prevent this

interaction were it ranked higher above those two constraints, as in L58. Thus, in L58 we see s2 surface faithfully to stress despite having a long vowel, reflective of the dominance of WSP and ID[Length]. We may compare the two rankings side-by-side to see this effect in greater detail:

(25) L32: MR >> **ID[Length]** >> {NoLong, **WSP**} >> **ID[Stress]** >> {ML, Oblig}

(26) L58: MR >> **ID[Stress]** >> {ML, Oblig} >> **ID[Length]** >> {NoLong, **WSP**}

The difference between them is in the dominance relation between ID[Stress], ID[Length], and WSP. In L32 the ID[Stress] is dominated by the other two. We accordingly see underlyingly long suffixes surface long and stressed, thus the neutralization of s2 and s4. In L58, ID[Stress] dominates ID[Length] and WSP. Thus, s2 emerges faithful to stress and also length in L58. The position of ID[Stress] relative to WSP tracks restrictiveness: in the more restrictive grammar (L32), it is below WSP, while in the less restrictive grammar (L58) it is above it.

2.5 Case 2: a Phonotactic Identity

2.5.1 Language L45, the subset language

Language L45 neutralizes length on unstressed vowels. Stress surfaces faithfully when only one morpheme is stressed underlyingly, but if neither morpheme is specified for stress, the word will not have surface stress. Stress defaults to the suffix when both morphemes are underlyingly stressed. However, in that case if one morpheme is long and stressed, it will surface stressed. The description of Language L45 is presented in (27) along with a grammar that generates it:

(27) a. WSP >> ID[Stress] >> Oblig >> ID[Length] >> {NoLong, MR} >> ML

b. Language L45

	r1 = /pa/	r2 = /pa:/	r3 = /pá/	r4 = /pá:/
s1 = /ka/	paka	paka	páka	pá:ka
s2 = /ka:/	paka	paka	páka	pá:ka
s3 = /ká/	paká	paká	paká	pá:ka
s4 = /ká:/	paká:	paká:	paká:	paká:

In L45, there are total neutralizations of r1/r2, and s1/s2. Both r2 and s2 are long underlyingly, but both will surface short and unstressed because ID[Stress] and WSP are high ranked: WSP neutralizes faithfulness to length in favor of faithfulness to stress.

We can thus illustrate morpheme behavior more clearly in the following chart:

(28) L45 Compressed

	/pa/, /pa:/	/pá/	/pá:/
/-ka/, /-ka:/	paka	páka	pá:ka
/-ká/	paká	paká	pá:ka
/-ká:/	paká:	paká:	paká:

The phonotactic inventory of L45 has five words:

(29) L45 Phonotactic Inventory: paka, páka, pá:ka, paká, paká:

2.5.2 Language L25, the superset language

Language L25 neutralizes length on unstressed vowels. If a morpheme is long underlyingly, it will surface stressed, but when forced to choose between initial or final stress, stress will fall finally. A ranking generating L25 is shown in (30a), and the language itself is shown in (30b).

(30) a. WSP >> ID[Length] >> NoLong >> ID[Stress] >> {MR, Oblig} >> ML
b. Language L25

	r1 = /pa/	r2 = /pa:/	r3 = /pá/	r4 = /pá:/
s1 = /ka/	paka	pá:ka	páka	pá:ka
s2 = /ka:/	paká:	paká:	paká:	pá:ka
s3 = /ká/	paká	pá:ka	paká	pá:ka
s4 = /ká:/	paká:	paká:	paká:	paká:

There are no total neutralizations in L25: each pair of roots and each pair of suffixes behave differently in at least one environment.

The phonotactic inventory of L25 has five words:

(31) L25 Phonotactic Inventory: paka, páka, pá:ka, paká, paká:

2.5.3 L45 is a paradigmatic subset of L25

Unlike the previous case of L32 and L58, L45 and L25 are phonotactically identical: both languages neutralize every possible input to the same five surface forms: paka, páka, pá:ka, paká, paká:. However, while L25 contains no total neutralizations, L45 neutralizes both s1/s2 and r1/r2. If we compare the morphemic behaviors of the two languages side-by-side, we see that L45 forms a paradigmatic subset of L25.

(32) L45 Compressed

	/pa/, /pa:/	/pá/	/pá:/
/-ka/, /-ka:/	paka	páka	pá:ka
/-ká/	paká	paká	pá:ka
/-ká:/	paká:	paká:	paká:

(33) L25 with L45-equivalent shaded subset

	r1 = /pa/	r2 = /pa:/	r3 = /pá/	r4 = /pá:/
s1 = /ka/	paka	pá:ka	páka	pá:ka
s2 = /ka:/	paká:	paká:	paká:	pá:ka
s3 = /ká/	paká	pá:ka	paká	pá:ka
s4 = /ká:/	paká:	paká:	paká:	paká:

The morphemic behavior in L45 is a subset of the morphemic behavior of L25, as represented by the shaded cells in (33). L25 contains four root behaviors and four suffix behaviors, while L45 has three of those root and three of those suffix behaviors. Unlike with languages L8 and L7 in the SL system, there is only one way to "project" L45 into L25.

If we examine rankings generating the two languages side by side, we see why in L25 r2 and s2 are not neutralized: ID[Length] dominates ID[Stress], thus each morpheme will surface faithful

to length, and stressed because of the high position of WSP. The crucial difference between the two languages then lies in the relative dominance between the two faithfulness constraints. If faithfulness to stress is higher, unstressed morphemes will surface unstressed and WSP will force unfaithfulness to length. If faithfulness to length dominates, we will see unstressed long morphemes surface long and stressed.

(34) L45: WSP >> **ID[Stress]** >> Oblig >> ID[Length] >> {NoLong, MR} >> ML

(35) L25: WSP >> ID[Length] >> NoLong >> **ID[Stress]** >> {MR, Oblig} >> ML

The position of ID[Length] correlates with restrictiveness: in the more restrictive grammar (L45) is is lower, but in the less restrictive grammar (L25) it is higher.

2.6 Discussion

The SL system contained a case of paradigmatic subsets. L8, the subset language, contained morphemic neutralization of stress in both roots, and suffixes. L7, the superset language contained no morphemic neutralization. We furthermore saw that the subset relation between the two languages changed depending on the underlying value assigned to neutralizing morphemes of L8. Importantly, when the morphemes of L8 were either all stressed or all unstressed, though they projected into different phonotactic forms in L7, those phonotactic forms were nonetheless forms appearing in L8.

But there was a third subset relation which yielded a different mapping that was inconsistent with the grammar of L8. That relation depended on the neutralizing morphemes of L8 having underlying feature values that were not the same, i.e. though r2 /pa:/ and r4 /pá:/ neutralize stress in L8, r2s3 /pa:ká/ maps to [paká] in L7, but in L8 it maps to [pá:ka]. The value assigned to a neutralizing feature in L8 made a difference to the surface realization of morphemes in L7. Crucailly, to achieve this correct mapping, two features values need to be set in one of the alternative morphemes.

As with the SL system, NSL contains paradigmatic subsets. However, the multiple projection feature of SL paradigmatic subsets did not hold for NSL subsets. For both L32 and L45, there was only one way to project into the superset language, regardless of the feature value assigned to the neutralizing features. L32 was also a phonotactic subset of L58, but L45 was phonotactically identical to L25. This difference will play an important role in maintaining restrictiveness, to be discussed in Section 3.3.2.

The next section introduces the Ouput-Driven Learner (ODL), the algorithm used to run learning simulations in this paper and in Tesar (2014).

3 Output-driven learning

The challenge of learning underlying representations in a phonological system lies in simultaneously learning those underlying forms from the observed ones, and the constraint hierarchy which determines the mapping between them. These two aspects are mutually dependent. Prior work on this topic includes particularly Merchant's (2008) Contrast Pairs and Ranking algorithm, Jarosz's (2006) Maximum Likelihood Learning of Lexicons and Grammars, Apoussidou's (2007) variation on the Gradual Learning Algorithm (Boersma 2001), and Tesar's (2014) Output-Driven Learning algorithm (ODL).

The Ouput-Driven Learner incrementally learns both underlying forms and constraint rankings from observed data. It exploits logical structure in the space of possible input forms. This structure is output-drivenness, presented in Section 3.1. The learning algorithm itself is presented in Section 3.2.

3.1 Output-Driven-Maps

At the heart of the ODL algorithm are output driven maps. For any output, there is a similarity-ordering defined on the space of possible inputs. A candidate includes an input/output pair, the feature disparities between them, and the IO correspondence relation.

Assuming the features +/- stress and +/- length, the output [paká:] has the featural specification (-,-)(+,+). A possible input /páká:/ creates the candidate /páká:/[paká:] with a single featural disparity in stress on the first syllable. A second possible input /pá:ká:/ forms a second candidate with the same output [paká:]. That second candidate /pá:ká:/[paká:] has two disparities, both of stress and length in the first syllable. Both inputs share a feature disparity in stress with the output, but the latter has an additional disparity in length. The first candidate /páká:/[paká:] has greater *internal similarity* than the second candidate /pá:ká:/[paká:] exactly because the disparities between the input and the output in the first candidate are a subset of the disparities between the input and the output in the second candidate.

Relative similarity is a partially ordered relation between candidates, based on internal similarity. A similarity lattice is given in Figure 1 for output [paká:]. At the top of the lattice is the

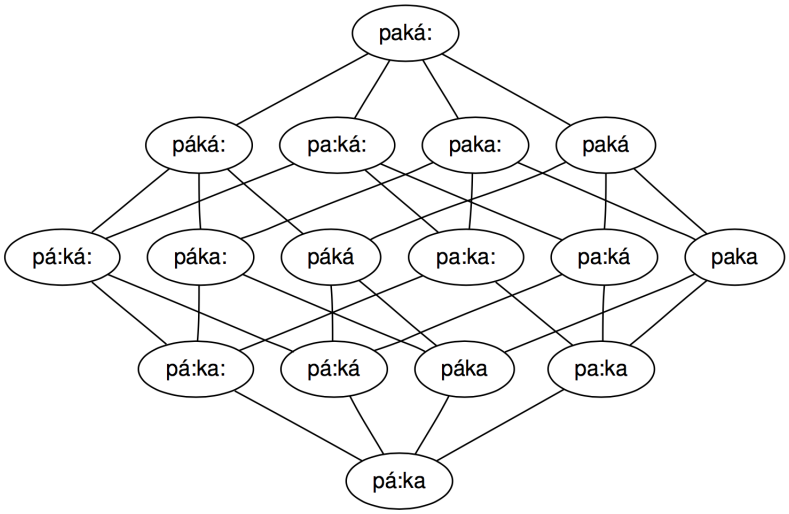

Figure 1: Relative Similarity Lattice for [paká:].

candidate with greatest internal similarity /paká:/[paká:], and at the bottom is the one with least internal similarity (/pá:ka/[paká:]. A map is *output-driven* if, for any input/output map, any other input with greater similarity to the output will also map to that output. The identity input (the one at the top of the lattice) will always map to the output, because the map is idempotent. If the /pá:ká:/ (the input on the far left of the third row) maps to [paká:], then so do /páká:/ and /pa:ká:/. This relationship can be thought of as one of entailment: since the form with one disparity has a subset of the disparities that the two-disparity form has, the former is entailed by the latter. Likewise, as the zero-disparity form at the top of the lattice has a subset of both, it is entailed by both.

3.2 The learning algorithm

The Output-Driven Learner must appropriately capture the morphemic alternations and the underlying forms of the morphemes, but must also capture restrictions that might be present even when there is no surface indication. The ODL algorithm encodes an incremental process that see-saws between these mutually-dependent aspects of the learning process. As it gains a little informa-

tion about the output forms, it uses that information to determine a little information about the ranking of some constraints, back and forth. Learning underlying forms and constraint rankings is enhanced by exploiting the output-driven maps associated with the phonotactic forms that the learner observes. If any input maps to the observed form, then the maximally similar input will also because of output-drivenness.

Phonotactic Learning takes advantage of this fact to gain initial information about the constraint ranking. The learner tests the consistency of each maximally-similar input to the observed form against an initial ranking that ranks faithfulness below markedness constraints. If the maximally similar input is not optimal for some ranking, i.e., there is a ranking that prefers a winner other than the target output form, then a winner-loser pair is formed. The learner does this for each of the observed forms, collecting a set of winner-loser pairs, called *the support*. It encodes facts about which rankings prefer the winner to the loser of a pair.

Once Phonotactic Learning has ceased, the learner moves on to Single Form Learning, when features for underlying forms may be set. This assumes that the learner has access to information about the relationships between different morphemes in the observed forms: what are the roots, suffixes and when are they the same, different or just alternating in certain environments. First for each word, it tests the maximally *dissimilar* input—if this candidate maps to the word's output for the generated ranking, then all other viable inputs will too. This is Initial Word Evaluation. If it is inconsistent, it moves on to test each single-disparity input against the support gained from phonotactic learning. If a single-disparity candidate is inconsistent with the support, the learner sets the inconsistent feature to its opposite value, which will always be the value of the feature in the corresponding output segment. For example, for the output form [paká:], if the single-disparity candidate /paká/ with the suffix length feature of -long is inconsistent, the underlying value for length in the suffix must be +long.

As the learner has access to paradigmatic information, at this point it looks to see whether the feature for that morpheme alternates values in any observed forms. If it does, the learner proceeds with testing the consistency of features of other words in which that morpheme appears. When the learner evaluating r1s4 [paká:] finds that the minimal disparity candidate /paká/ is inconsistent, it sets s4's length feature in the lexicon to +long underlying. With this feature set, several possible inputs are ruled out, as shown in Figure 2. Viable candidates are not shaded.

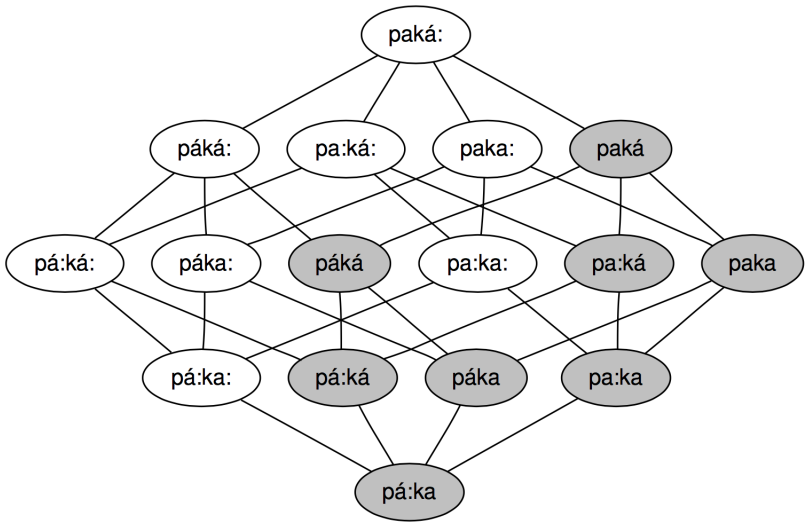

Figure 2: Viable Similarity Lattice for [paká:]. When the suffix is set to +long in the lexicon, all candidates with a -long suffix are ruled out on the basis of inconsistency.

Imagine that s4's length feature alternates on the surface: word r3s4 surfaces as [páka]. The learner constructs a candidate with all unset features faithful to the surface form, and with set features faithful to underlying form: candidate /páka:/[páka]. This candidate is evaluated with a generated test ranking, and if it is not optimal for that ranking, the candidate is adopted as the winner in a winner-loser pair, and that winner-loser pair is added to the support as a piece of non-phonotactic ranking information.

The third phase of learning occurs when no further features can be set on the basis of single form learning alone, but some forms are still failing initial word evaluation (testing the maximal disparity candidate). The learner's solution is to process two words at the same time, a contrast pair, which provides certain crucial pieces of information.

Learning could finish at any point, as soon as all words pass initial word evaluation. Furthermore, learning can be successful without all underlying feature values having been set as long as surface contrasts are accounted for. Any features left unset after successful learning are necessarily non-contrastive in all environments, and for a given combination of underlying forms for a word, any assignment of values to unset features will map to the same output. The process is iterative and relatively fast. The next section discuss the ODL's measures for enforcing restrictiveness.

3.3 Existing restrictiveness methods in the ODL

Restrictiveness in Tesar (2014)'s ODL is enforced through two methods which differ in the location of the enforcement. The first method involves the generation of a ranking from the support. The second involves setting features in the lexicon. These two existing methods will be discussed in the next two sections.

3.3.1 Restrictiveness in the ranking

Biased Constraint Demotion (henceforth BCD, Prince & Tesar 2004) is a method for enforcing restrictiveness in a grammar without the learner having to compute all pairwise restrictiveness comparisons between grammars in the typology. It is a recursive algorithm that ranks constraints in a stratified hierarchy to generate a ranking out of the support (a set of winner-loser pairs). Winner-loser pairs encode the preference of each constraint in the grammar for one or the other of the pair.

BCD ranks a constraint if it prefers only winners. Thus the first stratum of the hierarchy will contain only constraints which never prefer any loser in the collection of winner-loser pairs. When a constraint is ranked, the winner-loser pairs which account for its ranking are removed from the support. Both the constraint and the ERCs are dismissed from the remainder of the computation. The support thus shrinks over the course of computation until all winner-loser pairs are accounted for. If all constraints can be ranked, and all winner-loser pairs dismissed, a consistent ranking is generated. If at a point no more constraints may be ranked because none prefer only winners, and some winner-loser pairs remain unaccounted for, then no consistent ranking can be generated; the support is inconsistent.

Rather than rank all constraints preferring winners as high as possible, BCD can first rank markedness constraints before faithfulness constraints (a markedness-over-faithfulness, or FaithLow, bias) or faithfulness constraints before markedness constraints (a faithfulness-over-markedness, or MarkLow, bias). The choice of one bias over another will depend on the stage of learning. During Phonotactic Learning, the learner tests inputs which are fully faithful to observed outputs. Faithfulness constraints will naturally prefer those faithful winners, and no other competitors will do better with respect to faithfulness constraints. Without a bias, the faithfulness constraints would be ranked at the top of the stratified hierarchy. The learner then uses a mark-over-faith-edness bias.

By using such a bias where faithfulness constraints are at the bottom of the ranking by default, when the learner encounters a word that requires having a faithfulness constraint higher up,

it must construct a winner-loser pair which forces that faithfulness constraint to be higher, and add that pair to the support, thus obtaining specific information about the ranking.

Sometimes there may be no markedness constraint that fits the criterion of preferring no losers. This could mean one of two things. On the one hand, if none of the remaining constraints (markedness or faithfulness) prefers only winners, then no consistent ranking can be generated from the support: the support is inconsistent. On the other hand, if there is a faithfulness constraint which does prefer only winners, it may be ranked. When the ERC accounting for that faithfulness constraint is removed from the support, it may then be possible to rank markedness constraints if the dismissed ERC removes evidence that those constraints prefer losers. Indeed, this is the desired outcome of ranking a faithfulness constraint, only as needed to "free up" markedness constraints for ranking.

If there is more than one faithfulness constraint preferring no losers, BCD will choose the one that frees up the most markedness constraints for ranking. In other words, BCD chooses the faithfulness constraint that will dismiss ERCs with information about markedness constraints preferring losers. However, choosing a faithfulness constraint to rank is not always straightforward in particular if two faithfulness constraints free up the same number of markedness constraints.

Prince and Tesar (2004) proposed the r-measure as an easily computable approximation of the restrictiveness of a grammar. It is calculated by summing up the number of markedness constraints dominating each faithfulness constraint. The higher the r-measure, the more restrictive the grammar. For illustration, recall the two constraint rankings discussed in Section 2.1, which generate Languages L8 and L7, respectively:

(36) L8/L7 from SL:
   a. L8: WSP >> ID[Length] >> NoLong >> MR >> ML >> ID[Stress]
   b. L7: WSP >> ID[Stress] >> ID[Length] >> NoLong >> MR >> ML

Recall that L8 is a paradigmatic subset of L7; its more restricted distribution of stress makes it the more restrictive grammar. The r-measure accurately tracks restrictiveness between these two grammars: for L8 it is 5 (1+4) while for L7 it is 2 (1+1).

In most cases, the r-measure satisfactorily tracks the restrictiveness of a grammar. However, it is not full-proof. Let us now turn to the two cases we saw in NSL, Section 2.3:

(37) L32/L58 (Case 1) from NSL:
   a. L32: MR >> ID[Length] >> {NoLong, WSP} >> ID[Stress] >> {ML, Oblig}
   b. L58: MR >> ID[Stress] >> {ML, Oblig} >> ID[Length] >> {NoLong, WSP}

(38) L45/L25 (Case 2) from NSL:
   a. L45: WSP >> ID[Stress] >> Oblig >> ID[Length] >> {NoLong, MR} >> ML
   b. L25: WSP >> ID[Length] >> NoLong >> ID[Stress] >> {MR, Oblig} >> ML

L32 and L58 have an r-measure of 4 (1+3), while L45 and L25 both have an r-measure of 2 (1+1), but L32 is more restrictive than L58, and L45 is more restrictive than L25. In the pair L32 and L58, BCD has to choose between ranking ID[Length] first to free up NoLong and WSP, or ID[Stress] to free up ML and Oblig. The former choice will lead to a more restrictive grammar. In the pair of L45 and L25, BCD can rank ID[Stress] to free up Oblig, or ID[Length] to free up NoLong, but only the former choice leads to a more restrictive grammar.

BCD approximates the r-measure by its very nature: it maximizes the domination of faithfulness by markedness constraints. However, neither the r-measure nor BCD can distinguish the more restrictive grammar in these two cases from NSL because the more restrictive grammar is determined by the choice between ranking one faithfulness constraint before the other. These constraints are treated only in terms of the number of markedness constraints freed up, and they free up the same number. Thus there is no way for BCD to determine the more restrictive ranking with

respect to ranking one faithfulness constraint over the other. Prince and Tesar (2004) discuss several other cases which illustrate the limitations of the r-measure.

3.3.2 Restrictiveness in the lexicon

Fewest Set Features is a method for enforcing restrictiveness through feature setting in the lexicon. Restrictiveness in the lexicon can be viewed in terms of mapping more inputs to fewer outputs: the number of set features in the lexicon negatively correlates with a grammar's degree of restrictedness because it decreases the amount of inputs which neutralize to surface forms, and decreases restrictiveness. The learner should then only set features as necessary to preserve as much neutralization as possible while accounting for all the surface contrasts. Minimizing the number of set features can work if the learner is committed to setting only features that must be set to a specific value, to account for phonotactic contrasts. Though the ODL generally embodies this feature setting conservatism, Fewest Set Features embodies it explicitly when BCD alone is insufficient to discern the most restrictive ranking because the learner lacks necessary information about the underlying feature values in the lexicon.

In Language L8 presented in Section 3, the learner cannot set the length feature of r1 because of the multiple paradigmatic subset relations between L8 and L7. During Initial Word Evaluation, it first tests the form with maximally mismatched features at the bottom of the similarity lattice. If this word fails evaluation, the learner knows that more information is needed, either ranking information is incomplete or some unset features in that word must be set. For L8, word r1s1 [paká] does not pass. As s1 is set in the lexicon to -long, the maximally mismatch input is /pá:ka/.

Recall from Section 2.1.3 the multiple ways to "project" L8 into L7. The first two projections set the neutralizing morpheme r1 to either + or - long, and the target grammar of L8 mapped these correctly to the target surface form [paká]. The third projection required r1 be specified for +long and -stress, but the target grammar of L8 did not map that morpheme to the correct output. The learner's support at the time when learning halts is consistent with grammars generating both L7 and L8, and therefore either feature setting route will be consistent with the support.

Fewest Set Features steps in. Using a word which fails Initial Word Evaluation, the learner tests each of the unset features in an analogous way to feature testing during Single Form Learning: for each unset feature, it constructs an input matching the surface realization for that feature, and the remaining unset features mismatching their surface values. These are inputs one row up from the bottom row. Figure 3 shows the viable similarity lattice for r1s1.

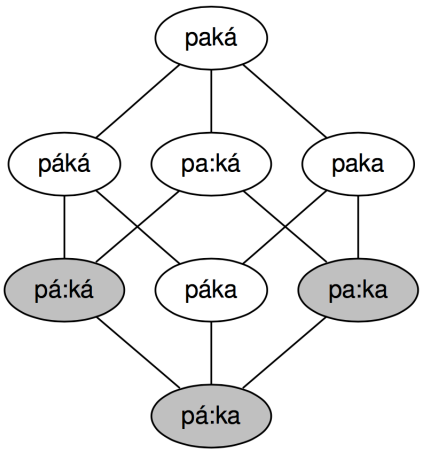

Figure 3: Viable Similarity Lattice for r1s1 [paká]. Shaded inputs are inconsistent.

There are three viable inputs on that row: pá:ká, páka and pa:ka. Each candidate is then evaluated for consistency with the support, in combination with candidates for each of the words passing initial word evaluation. If inconsistency results, then the learner knows that the evaluated input for r1s1 is inadequate.

If there is a ranking which renders the candidate for r1s1 plus all of the words passing Initial Word Evaluation, then the option of setting the tested feature may be pursued. Given output-drivenness, all words above that bottom candidate in the relative similarity lattice will be consistent. If none of the candidates on the second row up pass, then the learner continues up another row, testing inputs until a passing one is found and a feature can be set. Once a feature is set, the learner pursues learning by looking for surface alternations and non-phonotactic ranking information.

For r1s1 in L8, /páka/ is the only input from the second-to-bottom row that yields a consistent mapping, thus the other forms are shaded. However, there is another potential mapping with an input one row higher, an input that is not ordered with respect to /páka/; this input is /pa:ká/. These two candidates, /páka/[paká] and /pa:ká/[paká] are two possible projections of L8 into L7 as discussed above. The first input requires only one feature to be set, while the second input requires two features to be set. The restrictiveness enforcing Fewest Set Features will choose the input that requires the least number of features to be set. This is the input on the second-to-bottom row, /páka/. The learner will then choose that as the input to word r1s1 [paká], the choice yielding the more restrictive grammar.

4 Results of learning

This section presents the results of the learning simulation for both SL and NSL. All 24 languages of SL were successfully learned using the ODL's current formalization discussed above, including two paradigmatic subsets, L8 and L17. Section 4.1 briefly discusses the simulation of SL, and in particular learning L8. L17 will not be discussed because it is a symmetric version of L8.

The results of learning in SL are presented for comparison with the results of learning in NSL because the existing restrictiveness measures discussed in Section 3.3 are sufficient for learning SL, but insufficient for learning NSL. Nonetheless, 58 out of the 62 languages of NSL were also successfully learned. The four unsuccessfully learned languages reveal themselves to be paradigmatic subsets of four successfully learned ones, two of which were presented in Sections 2.4 and 2.5. The languages not discussed in this paper are symmetric versions of the languages that are discussed. Section 4.2 discusses the learning simulation NSL, in particular Language L32 in Section 5, and Language L45 in Section 6.

4.1 Learning SL

The SL system has 24 languages. The learning simulation was successful for all 24 of those languages, including the cases of paradigmatic subsets, L8 and L17. Both languages were successfully learned with the addition of the Fewest Set Features method. The paradigmatic properties of L8 were presented in Section 2.1, and this section will discuss the results of learning for that language.

The support accumulated by the learner is in (39). BCD generates rankings consistent with both L8 and L7, even using the restrictiveness enforcing FaithLow bias. BCD fails in this case because underlying features which are not yet set are required to account for surface contrasts, in particular the length feature of r1 cannot be set with that generated ranking.

(39) Learned ranking information of L8 (incomplete):

Word	Input	Winner	Loser	WSP	ID[L]	NoLong	MR	ML	ID[S]
r2s2	pa:ká:	paká:	pa:ká:	W	L	W			
r1s2	paká:	paká:	paká		W	L			
r2s1	pá:ka	pá:ka	paká		W	L	L	W	W
r1s1	paká	paká	páka				W	L	W

(40) Learned lexicon for L8 (incomplete)

r1	/?,?/	r2	/?,+/
s1	/?,-/	s2	/?,+/

The information above represents the learner's hypothesis and lexicon before applying Fewest Set Features. As discussed in Section 3.3.2, word r1s1 failed Initial Word Evaluation. The learner tested each of the viable inputs on the second row from the bottom, /pá:ká/, /páka/, and /pa:ka/. Only one of those inputs passed Fewest Set Features: /páka/. The learner was able to set r1 to -long underlyingly. The resultant updated lexicon is below in (41):

(41) Updated and completed lexicon for L8

r1	/?,-/	r2	/?,+/
s1	/?,-/	s2	/?,+/

With the new lexicon in hand, the learning of L8 is able to successfully complete. The ranking information in (39) is sufficient, and all words pass Initial Word Evaluation.

4.2 Learning NSL

Of the 62 total languages in NSL, 58 were successfully learned. The failed languages are paradigmatic subsets of languages which are learned successfully. These four languages are listed below, along with rankings that generate them.

(42) a. MR >> ID[Length] >> {NoLong, WSP} >> ID[Stress] >> {ML, Oblig}

b. Language L32

	r1 = /pa/	r2 = /pa:/	r3 = /pá/	r4 = /pá:/
s1 = /ka/	paka	pa:ka	paka	pa:ka
s2 = /ka:/	paká:	pa:ká:	paká:	pa:ká:
s3 = /ká/	paká	pa:ká	paká	pa:ká
s4 = /ká:/	paká:	pa:ká:	paká:	pa:ká:

(43) a. WSP >> ID[Stress] >> Oblig >> ID[Length] >> {NoLong, MR} >> ML
b. Language L45

	r1 = /pa/	r2 = /pa:/	r3 = /pá/	r4 = /pá:/
s1 = /ka/	paka	paka	páka	pá:ka
s2 = /ka:/	paka	paka	páka	pá:ka
s3 = /ká/	paká	paká	paká	pá:ka
s4 = /ká:/	paká:	paká:	paká:	paká:

(44) a. WSP >> ID[Stress] >> Oblig >> ID[Length] >> {NoLong, ML} >> MR
b. Language L46

	r1 = /pa/	r2 = /pa:/	r3 = /pá/	r4 = /pá:/
s1 = /ka/	paka	paka	páka	pá:ka
s2 = /ka:/	paka	paka	páka	pá:ka.
s3 = /ká/	paká	paká	páka	pá:ka
s4 = /ká:/	paká:	paká:	paká:	pá:ka

(45) a. ML >> ID[Length] >> {NoLong,WSP} >> ID[Stress] >> {MR, Oblig}
b. Language L59

	r1 = /pa/	r2 = /pa:/	r3 = /pá/	r4 = /pá:/
s1 = /ka/	paka	pá:ka	páka	pá:ka
s2 = /ka:/	paka:	pá:ka:	páka:	pá:ka:
s3 = /ká/	paka	pá:ka	páka	pá:ka
s4 = /ká:/	paka:	pá:ka:	páka:	pá:ka:

Examining these more closely, a pattern emerges: L32 and L59, and L45 and L46 are mirror images. The only difference between each pair is the relative ranking of ML and MR. In L32, MR is ranked highest, while ML is on the lowest stratum with Oblig; L59 shows the opposite ranking: ML is high and MR is low. In L45, MR >> ML, but in L46 ML >> MR.

Based on these observations, we note that there are in fact two cases where learning fails. The two cases of this differ in one crucial way: one language forms not only a paradigmatic, but also a phonotactic subset of the successful language. In the other case, the language forms a paradigmatic subset and a phonotactic identity with a successfully learned language. The remainder of this paper will focus on L32 and L45 as representatives of the above two cases, respectively: Section 5 discusses the learning simulation in L32, and Section 6 discusses the results of L45.

4.3 Discussion

The restrictiveness measures as currently formalized in the ODL were sufficient for the successful learning of the SL typology, and although four languages in NSL failed, 58 languages were successful with current restrictiveness measures. 24 of those are the same as in SL because Oblig was ranked high enough to block stressless outputs from surfacing. The remaining 38 languages contained stressless outputs. A closer examination of those reveals that 17 languages implemented the Fewest Set Features to successful end: L26, L28, L31, L33, L35, L36, L37, L38, L39, L45, L46, L48, L49, L50, L52, and L61. Though L45 and L46 were ultimately not learned, Fewest Set Features correctly set some, though not all, unset features. However, in those two languages as well as a couple others, the method incorrectly set some features which did not contrast on the surface. This result will be discussed in greater detail in Section (6).

5 Language L32, a phonotactic subset of L58

5.1 Learning Fails

The ODL applied to L32 is successfully able to set all surface contrasting feature in the lexicon, as presented in (47). However, learning cannot proceed further than Single Form Learning. The resulting support is presented below in (46).

(46) Learned ranking information of L32 (incomplete):

Word	Input	Winner	Loser	MR	ID[S]	ML	Oblig	ID[L]	NoLong	WSP
r1s1	paka	paka	paká		W	W	L			
r1s2	paká:	paká:	paka		W	L	W	W	L	
r1s3	paká	paká	paka		W	L	W			
r1s2	paká:	paká:	paká					W	L	
r2s1	pa:ka	pa:ka	paka					W	L	L

(47) Learned lexicon for L32 (complete)

r1	/?,-/	r2	/?,+/	r3	/?,-/	r4	/?,+/
s1	/-,-/	s2	/?,+/	s3	/+,-/	s4	/?,+/

The lexicon is complete. Because L32 has no morphemic alternations, no non-phonotactic ranking information can be found during feature setting. Nor can any contrast pair can then be formed. Furthermore, the features set in L32 are a subset of the features set in L58.

Word r1s2 [paká:] fails Initial Word Evaluation. The viable sublattice for [paká:] is shown in Figure 4: r1 and s2 are both unset for stress, but they do not contrast in stress and so do not need to

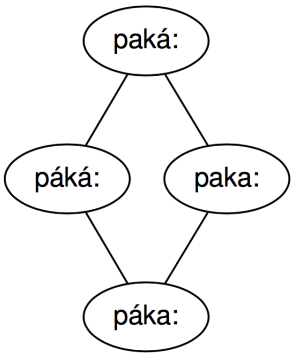

Figure 4: Relative Similarity Lattice for r1s2 [paká:] in L32.

be set in the lexicon. However, the learner attempts to set those unset features by applying Fewest Set Features. The word on the bottom of the viable similarity lattice is input /páka:/. However, when the learner tests the two viable inputs on the row second from the bottom, testing r1 for -stress in /páká:/ and s2 for +stress in /paka:/, it finds that both mappings yield consistency with the current support in (46). Rather than choosing at random and setting one of the two features, the learner declares learning unsuccessful.

On a deeper glance, this makes sense because at the end of Single Form Learning, the lexicon is complete. Furthermore, L32 has no morphemic neutralization from which non-phonotactic ranking information can be gained. Finally, the features set in L32 are a subset of the features set in L58. This means that any contrastive underlying feature in L32 will be consistent with L58, a final blow against a feature setting approach to restrictiveness.

Using a Mark-over-Faith bias, BCD will rank MR on the first stratum as it prefers no losers. Either faithfulness constraint will free up two markedness constraints for ranking: ID[Stress] frees ML and Oblig; ID[Length] frees up NoLong and WSP. The support is consistent with both rankings. Ranking ID[Stress] first will yield a grammar consistent with L58, the superset, while ranking ID[Length] first yields a grammar consistent with L32. Further, these two grammars tie for the best r-measure (4).

(48) a. L32: MR >> ID[Length] >> {NoLong, WSP} >> ID[Stress] >> {ML, Oblig }
b. L58: MR >> ID[Stress] >> {ML, Oblig} >> ID[Length] >> {NoLong, WSP}

What is crucially missing from the support in (46), is that WSP dominates ID[Stress] in L32 but not in L58. This is what causes the neutralization of stress across s2 /-,+/ and s4 /+,+/ in L32, but not in L58.

5.2 A proposed solution: getting a hold of implicitly represented non-phonotactic information

The ODL only commits to a hypothesis insofar as evidence can be fully justified on the basis of the observed forms of the language. This evidence comes from error detection during the phonotactic stage of learning, when errors reveal ranking information necessary to enforce the fully faithful mappings (idempotency). Thus, the learner works with the observed phonotactic forms to gain ranking information on the basis of the fully faithful or minimal disparity inputs. The winner-loser pairs collected from Phonotactic Learning will then only reveal explicit ranking information about what cannot be kept from surfacing.

If a language lacks certain forms, the Output-Driven Learner will not test those forms during Phonotactic Learning (though possibly in Single Form Learning). Since faithfulness constraints encode information about disparities between input and output mappings, phonotactically illicit inputs will violate faithfulness constraints because they map to outputs that are phonotactically viable, and thus unfaithful. Recall that the missing ranking information for L32 boiled down to the dominance of ID[Stress] by WSP. A winner-loser pair that encodes this information should contain an input that violates ID[Stress]. Let us re-examine the phonotactic words of L32, and the inputs which must neutralize to those words:

(49) *L32 Phonotactic Inventory*: paka, pa:ka, paká:, pa:ká:, paká, pa:ká

(50) *Non-phonotactically viable inputs for L32*:
a. paka:, pa:ka:
b. páka:, páka, pá:ka, pá:ka:
c. páká, pá:ká, páká:, pá:ká:

In fact, the learner's support in (46) only contains winner-loser pairs with ranking information from Phonotactic Learning only; winners are only the fully faithful forms in (49). Thus ID[Stress] will always prefers the winner, though the information that ID[Stress] must be dominated by WSP is the key difference between L32 and L58.

Since inputs in (50) must neutralize to the outputs in (49), in principle the learner can perform error detection and test for inconsistency for mappings between the phonotactically viable words and each illicit word. If, for a given input in (50), when it is paired with each of the words in (49), there is only one mapping consistent with the learner's support when learning fails, then the learner can know that this input is correct for the phonotactic word that forms a consistent

mapping. Furthermore, if a winner-loser pair is formed, that ERC can be added to the support as a piece of non-phonotactic ranking information.

So, there are two parts of this proposal which must be articulated. The first question is which of the forms in (50) might the learner test first? GEN rbans the multiple stress forms in (50c) from surfacing in any language. L58 contains the forms in (50a), which the learner's current support also allows. Neither L32 nor L58 contain the forms in (50b). However, the distinction between the forms in (50a) and (50b) is not directly accessible to the learner.

We want to see if any evidence would be available to the learner about the dominance of ID[Stress] by WSP. Intuitively, there seems to be a salient phonotactic generalization in (49): L32 only allows stress to be final. A plausible starting place is then is to choose an input that seems to violate this generalization. An initial stress input mapping to a final stress output will necessarily receive a violation from ID[Stress]. Recall that neither ID[Stress] nor ID[Length] in L32's support in (46) prefer the loser in any of the winner-loser pairs. If we can find a winner-loser pair where ID[Stress] prefers the loser, then BCD will be forced to rank ID[Length] instead of ID[Stress].

The choice to test an initial stress input is based on an intuitive observation about the phonotactics of L32. Future work will need to have more robust criteria for evaluating which generalization *the learner* might be able to exploit, and then which input it should choose to test on the basis of that generalization. In fact, many different phonotactic generalization might in principle be available for exploitation in this solution. Future work must characterize a principled way to decide which forms to test.

Second, the learner only commits to a grammar insofar as the evidence it has accumulated constitutes the "best guess" hypothesis given the evidence it has accumulated. The Output-Driven Learner (in fact, most learning algorithms) works on the basis of evidence as it comes in, and is designed to allow new evidence to dynamically effect the learner's hypothesis about the grammar. Simply, the door to learning is left open to allow for more data to be observed, no matter how infrequent that data might occur.

However, in order to deduce that a single consistent mapping between an input in (50) and an output in (49) is in fact correct, the learner must make an inductive leap. That leap is to assume that the phonotactic forms in (49) are indeed all and only the words of L32. If this leap can be made, it then follows that the words in (50) must map to the ones in (49). The second question is then when can the learner determine that it has seen enough data from a language to work confidently with both the phonotactics and the unobserved words in the manner described? The answer to this question is left to future research.

The next section discusses the results of this solution, which centers around the observation of distributional patterns within the data that the learner observes.

5.3 Results

The previous section concluded that the L32 phonotactic inventory contained a salient phonotactic pattern: no words have initial stress. If the learner can assume that it has observed all the words of L32, then it can concretely exploit this generalization and richness of the base to test the mapping between inputs which violate the generalization and the phonotactic forms which verify it.

As a first stab, the learner tests an input with initial stress, which violates the salient phonotactic generalization we identified. Incidentally, initial stress words are lacking from the phonotactic inventories of both L32 and L58. We chose input /páka:/ to test this solution. The learner's support at the time that learning fails, shown above in (46), generates a ranking consistent with L58, repeated below:

(51) MR >> ID[Stress] >> {ML, Oblig} >> ID[Length] >> {NoLong, WSP}

For input /páka:/, error detection reveals that [paka:] is optimal given the ranking generated. Note that this word is not in the phonotactic inventory of L32, but is in L58. Next, for each phonotactic

form in (49) above, six winner-loser pairs are formed, with input /páka:/, [paka:] adopted as loser, and each phonotactic form adopted as winner. This yields the six ERCs below in (52).

(52) Six mappings for input /páka:/

a. /páka:/[paka]

Input	Winner	Loser	MR	ID[S]	ML	Oblig	ID[L]	WSP	NoL
páka:	paka	paka:					L		

b. /páka:/[pa:ka]

Input	Winner	Loser	MR	ID[S]	ML	Oblig	ID[L]	WSP	NoL
páka:	pa:ka	paka:					L		

c. /páka:/[paká]

Input	Winner	Loser	MR	ID[S]	ML	Oblig	ID[L]	WSP	NoL
páka:	paká	paka:		L	L	W	L	W	

d. /páka:/[paká:]

Input	Winner	Loser	MR	ID[S]	ML	Oblig	ID[L]	WSP	NoL
páka:	pa:ká:	paka:		L	L	W	L		L

e. /páka:/[pa:ká]

Input	Winner	Loser	MR	ID[S]	ML	Oblig	ID[L]	WSP	NoL
páka:	pa:ká	paka:		L	L	W	L		

f. /páka:/[paká:]

Input	Winner	Loser	MR	ID[S]	ML	Oblig	ID[L]	WSP	NoL
páka:	paká:	paka:		L	L	W		W	

Each of these ERCs are in turn paired with the support in (46) for inconsistency detection. ERCs (52a) and (52b) are inconsistent on their own and ERCs (52c), (52d) and (52e) are inconsistent with the support.

The ERC in (52f) is consistent. The learner can permanently add this ERC as a piece of non-phonotactic ranking information to the support as in (53):

(53) Updated support for L32:

Word	Input	Winner	Loser	MR	ID[S]	ML	Oblig	ID[L]	NoLong	WSP
r1s1	paka	paka	paká		W	W	L			
r1s2	saká:	paká:	paka		W	L	W	W	L	
r1s3	paká	paká	paka		W	L	W			
r1s2	paká:	paká:	paká					W	L	
r2s1	pa:ka	pa:ka	paka					W	L	L
r1/r3-s2/s4	páka:	paká:	paka:		L	L	W		W	

Recall that previously, BCD had the choice to rank either ID[Stress] or ID[Length] because neither preferred a loser, and they freed up the same number of markedness constraints for ranking. This last ERC contains information that ID[Stress] prefers the loser to the winner, therefore ID[Stress] can no longer be ranked. This leaves ID[Length] free for ranking.

Using a Mark-over-Faith bias, the support now generates a ranking consistent with L32, and not L58:

(54) MR >> ID[Length] >> {NoLong, WSP} >> ID[Stress] >> {ML, Oblig }

This is indeed the target ranking for L32. The learner now has the missing information which distinguishes L32 from L58. Though this solution has not been fully implemented in the code, we may be confident that learning will now suceed for L32 with this updated support as it correctly generates a target ranking for L32.

5.4 Discussion

L32 contained a phonotactic generalization which could be exploited in order to make an inference about the neutralization of phonotactically illicit inputs. The generalization could only be made on the assumption that the learner has seen all the data of L32. Furthermore, the learner has a small set of unobserved forms which could easily be tested for consistency. However, these two assumptions could have limitations.

In the first place, a hypothesized neutralizing input may actually map to a form that is as yet unobserved but in the language. This is the risk of the inductive leap. The learner would then have erroneously mapped a neutralizing form onto an incorrect output. On the other hand, if the learner finds that a neutralizing input is inconsistent on every possible pairing with an observed form, this could indicate that more words need to be observed for that input.

In the second place, the space of possible inputs might be infinite. In the small systems defined here, the learner only has sixteen possible inputs to contend with, and in the case of L32 six phonotactic words. But modern theory assumes the space to be much larger, given the necessary features and combinatorial growth when all features are taken into consideration.

A related question arises, as to how the learner chooses inputs to test. As an initial criterion, we tested an input with initial stress because it seemed to violate a salient phonotactic generalization. The learner could have chosen any of the forms in (50). In fact, we can look closer at the results of different solutions. Both stressless inputs in (50a) provided a single consistent mapping to a surface form and a winner-loser pair to add to the support. Two of the four forms in (50b) did so as well, including the input discussed in the results. The four inputs in (50c), whose identity outputs GEN restricts, contained a single consistent mapping but provided no winner-loser pair to add to the support. In sum, out of the six inputs that might surface, four provided missing ranking information; a 2-in-3 chance of success. Since L32 differed in phonotactics from L58, a choice to test a phonotactically illict form provided missing ranking information.

Interestingly, the output which was optimal for the learner given the support at the time that learning fails (the error produced during error detection) and the chosen input /páka:/ was

not phonotactically viable in L32. Perhaps this observation could prove useful to the learner when choosing an input to test: any input mapping to a word that the learner has not observed might indicate to the learner that its current hypothesis is inadequate.

The proposed solution made concrete the possibility of using distributional evidence to boost learning. The fact that L32 formed a phonotactic subset of L58 perhaps allowed this solution to succeed. The next section discusses Language L45, where no distributional difference between the superset L25 exists.

6 Language L45, a phonotactic identity to L25

6.1 Learning fails

The ODL when encountering L45, constructs the following support (55) at the end of Single Form Learning, and the lexicon in (56). The information in both is incomplete. The ranking at the end of this stage of learning is consistent with the superset language L25.

(55) Learned ranking information for L45 (incomplete)

Word	Input	Winner	Loser	WSP	ID[L]	NoLong	ID[S]	MR	Oblig	ML
r4s3	pá:ká	pá:ka:	pa:ká	W				L		W
r4s4	pa:ká:	paká:	pa:ká:	W	L	W				
r1s4	paká:	paká:	paká		W	L				
r4s3	pá:ká	pá:ka	paká		W	L		L		W
r1s3	paká	paká	paka				W		W	L
r1s1	paka	paka	paká				W		L	W
r3s1	páka	páka	paka				W	L	W	
r3s3	páká	paká	páka					W		L

(56) Learned lexicon for L45 (incomplete)

r1	/-,?/	r2	/-,?/	r3	/+,-/	r4	/?,+/
s1	/-,?/	s2	/-,?/	s3	/+,-/	s4	/?,+/

The lexicon is incompletely learned in L45 at the end of Single Form Learning. The learner still must set the length features in r4 and s4, which are surface contrasting, and in principle should be able to be set. The learner can construct a contrast pair in attempt to set stress for r4, but the resultant ERCS are consistent with the support.

The learner applies the Fewest Set Features method four times, setting the stress features of r1, r2, s1 and s2.

(57) Lexicon learned after the fourth round of Fewest Set Features

r1	/-,-/	r2	/-,-/	r3	/+,-/	r4	/?,+/
s1	/-,-/	s2	/-,-/	s3	/+,-/	s4	/?,+/

Note that the Fewest Set Features method is setting features for morphemes which do no contrast on the surface, r1, r2, s1, s2. Furthermore, it is incorrectly setting r2 and s2 to -long underlying.

Recall that r1/r2 and s1/s2 neutralize globally in L45. The two feature values that need to be set (stress for r4 and s4) are not set.

The learner attempts Fewest Set Features to set stress for r4 and s4 with word r4s4 [paká:], which fails Initial Word Evaluation. The viable sublattice is given in Figure 5. As noted, r4 and

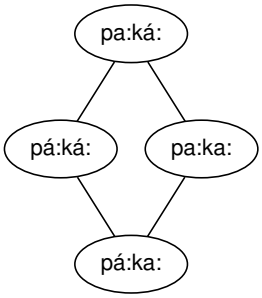

Figure 5: Relative Similarity Lattice for r4s4 [paká:] in L45.

s4 are both unset for stress but must be set to account for the surface contrast. Word r4s4 is thus a prime candidate for Fewest Set Features. The input at the bottom of the viable similarity lattice is /pá:ka:/. The learner tests the two unset features by temporarily setting r4 to -stress and s4 to +stress, yielding the two inputs on one row up from the bottom: /pa:ka:/ and /pá:ká:/, respectively. However, both inputs are consistent with the learner's current support. Rather than choose at random, the learner declares learning unsuccessful.

Using a Mark-over-Faith bias, BCD ranks WSP on the first stratum because it prefers no losers. This frees up either IDENT constraint to be ranked next, and one must be in order to free up markedness constraints. Ranking ID[Length] frees up NoLong, while ranking ID[Stress] frees up Oblig. Choosing ID[Length] yields a grammar consistent with L25, while choosing ID[Stress] yields one consistent with L45:

(58) L45: WSP >> ID[Stress] >> Oblig >> ID[Length] >> {NoLong, MR} >> ML

(59) L25: WSP >> ID[Length] >> NoLong >> ID[Stress] >> {MR, Oblig} >> ML

The support is consistent with both grammars. Note that the r-measure does not distinguish the two grammars: it is 3 for both. What is crucially missing is explicit information that ID[Stress] dominates ID[Length] in L45.

6.2 Discussion

Unfortunately, the solution proposed for L32 will not aid the learner in distinguishing L45 from L25. When phonotactically illicit inputs are paired with licit outputs, multiple ERCs are consistent with the support. The learner cannot be justified in adding any one ERC to the support.

Though more investigation is called for as to the reason our solution fails for L45, it is related to the fact that L45 and L25 are phonotactically identical. For L32, the phonotactically illicit input mapped to an output that was illicit for L32 but not L45. Thus, the solution succeeded on the basis of differing phonotactic distributions between L32 and L58. For L45, phonotactically illicit inputs will always map to allowed surface forms because L45 and L25 generate the same set of phonotactic forms. Thus, for a given input that must be neutralized, either both languages map it to the same output or they map it to two different outputs, both of which are phonotactically observed.

A successful solution for L45 will need to take advantage of the ways in which the paradigmatic behavior of L45 differs from L25. Since the lexicon and the ranking bias both fail to enforce restrictedness, the paradigmatic differences between the two languages might be exploited.

7 General Discussion and Future Directions

58 out of 62 languages of the NSL typology were successfully learned. The four failed languages reflect two cases of paradigmatic subsets. In both cases, the learner's support is consistent with a superset language. The two cases differ in phonotactic behavior with the superset: L32 is a phonotactic subset of L58, while L45 is phonotactically identical to L25.

The two languages differed in the course of learning as well. For L32, Single Form Learning learned the lexicon: all surface contrasts are accounted for in the lexicon by the end of that stage. For L45, all features but two are set, stress for r4 and s4, and length for s2 and r2 are incorrectly set to short using Fewest Set Features. It would seem that a lexical approach to restrictiveness whereby the learner attempts to set more underlying features is not useful, and is even leading the learner astray in L45. In both cases, Fewest Set Features fails to enforce restrictedness via the lexicon.

Biased Constraint Demotion also failed to distinguish the more restrictive ranking when multiple rankings are consistent with the support. In L32, the learner must see that WSP dominates ID[Stress]: in virtue of ID[Length]'s higher position in the ranking than ID[Stress], long suffixes will surface faithful to length. If they are long WSP will force them to be stressed regardless of underlying stress. In contrast, L58's grammar forces stress to surface faithfully regardless of underlying length because the ID[Stress] dominates ID[Length]. L45 also differs from the superset L25 in the relative ranking of ID[Length] and ID[Stress]: in L45 r1/r2 and s1/s2 neutralize to short because ID[Stress] dominates ID[Length]; but in L25, the reverse is true: s2 and r2 surface long and stressed.

It is *explicit* evidence about one faithfulness constraint preferring a loser that the learner needed to distinguish both subset from superset languages. For the learner's support in L32 and L45 when learning fails, BCD must choose a faithfulness constraint to rank, but neither constraint prefers a loser. Furthermore, both free up the same number of markedness constraints: the r-measure fails to distinguish the more restrictive ranking. This dilemma is classically noted in Prince and Tesar (2004). BCD approaches the ranking of faithfulness constraints only in terms of the number of markedness constraints that they free up when ranked. So, in our two cases, BCD treats them equivalently.

Where does this leave us? The learner crucially needs non-phonotactic ranking information to see evidence of neutralizations enforced by more restrictive rankings. But it cannot gain it using current strategies in the ODL. Both feature setting and a mark-over-faith-edness BCD fail to detect the more restrictive grammar.

Luckily, the phonotactic difference between L32 and L58 allowed successful completion of learning, if the learner could make an inductive leap and assume it had observed all the words of L32. Assuming this, a salient phonotactic generalization could be exploited to test an input that violates the generalization. The learner's support generated a ranking that did not map this input to a phonotactic word of L32, and the resultant pairing of the input with each phonotactic word of L32 yielded only one mapping consistent with the support. This mapping provided the learner with a crucial piece of non-phonotactic ranking information encoding the violation of ID[Stress] by the loser, which allowed BCD to generate a target ranking for L32.

The learner can infer correct ranking information for L32 because the space of possible outputs under consideration in this system is finite, and therefore computationally easy to test. As noted this solution will encounter problems when the space is infinite, and neutralized forms are not so easily induced. This is perhaps where a statistical approach can be incorporated to map a probability distribution over the space of possible inputs. More likely inputs would have a higher probability than less likely ones, and thus be better candidates for the learner to test. Future research can seek to incorporate such a statistical component.

The solution to L45 lies in exploiting the paradigmatic difference between it and L25 to acquire non-phonotactic ranking information. Such a distributional approach as was used for L32 will not be successful with L45 since its phonotactic forms are identical to those of L25. Further

research is needed to investigate a solution.

The NSL typology's modifications to SL allowed for words without main stress. A possible future direction might include expanding the system to model other empirical phenomenon, such as attested "pitch-accent" systems. In these languages such as Tokyo Japanese, accent marking appears to behave similarly to stress systems, while simultaneously allowing words without any accent. Though such phenomena are more diverse and (often) theoretically complex, future research might find idealizations appropriate to including them in the current model, insofar as they might be seen to share certain underlying properties and theoretical apparatuses.

What exactly is the relationship between non-obligatory stress and the failed languages? The addition of a single constraint and four more possible output forms exploded the typology to 62 languages. The first 24 languages of the NSL typology are identical to the languages in Tesar's typology because Oblig was ranked high enough to weed out stressless outputs. However, 38 languages did contain stressless outputs, and 17 of those required the Fewest Set Features method to complete learning. This fact indicates the presence of more paradigmatic subset relations than were explored in the current paper. The NSL typology itself warrants further investigation into general patterns in the typology, and a more thorough comparison with SL.

8 Conclusion

This paper presented an investigation of Output-Driven Learning in a new typological system modeling stress accent systems with both stressed and stressless words. By modifying GEN to allow these outputs without main stress, and adding a constraint which assigned violations to them, the system yielded a total of 62 languages. When the Output-Driven Learner was applied the typology, all languages were learned successfully except for four: these boiled down to two mirror-image cases exemplifying paradigmatic subsets. A solution was proposed where the learner constructs winner-loser pairs based on induced neutralizations. This solution successfully completed learning because the two languages contained phonotactic differences. The second case appears to be a vicious case of paradigmatic subsethood. Its solution is left for future research.

References

Akers, C. G. (2012). *Commitment-based learning of hidden linguistic structures*. PhD thesis, Rutgers University-Graduate School-New Brunswick.

Apoussidou, D. (2007). *The learnability of metrical phonology*. PhD thesis, University of Amsterdam.

Athanasopoulou, A., V. I. . P. N. An acoustic investigation of prosodic prominence in indonesian. (submitted).

Bernhardt, B. H. and Stemberger, J. P. (1998). *Handbook of phonological development from the perspective of constraint-based nonlinear phonology*. Academic press.

De Lacy, P. V. (2002). The interaction of tone and stress in OT. *Phonology*, 19:1–32.

Demuth, K. (1995). Markedness and the development of prosodic structure. *GLSA (Graduate Linguistic Student Association), Dept. of Linguistics, University of Massachusetts*.

Gnanadesikan, A. (2004). Markedness and faithfulness constraints in child phonology. *Constraints in phonological acquisition*, pages 73–108.

Gold, E. M. (1967). Language identification in the limit. *Information and control*, 10(5):447–474.

Hyman, L. M. (2009). How (not) to do phonological typology: the case of pitch-accent. *Language Sciences*, 31(2):213–238.

Jarosz, G. (2006). *Rich lexicons and restrictive grammars – maximum likelihood learning in Optimality Theory*. PhD thesis, The Johns Hopkins University, Baltimore, MD.

Kubozono, H. (2008). Japanese accent. In Miyagawa, S., editor, *The Oxford Handbook of Japanese Linguistics*. Oxford University Press.

Levelt, C. (1995). Unfaithful kids: Place of articulation patterns in early child language. talk. *Cognitive Science Department, The Johns Hopkins University, Baltimore, MD*.

McCarthy, J. J. and Prince, A. (1993). Generalized alignment. In Booij, G. and van Marle, J., editors, *Yearbook of Morphology*, pages 79–153. Dordrecht: Kluwer.

Merchant, N. (2008). *Discovering underlying forms: Contrast pairs and ranking*. PhD thesis, Rutgers University, New Brunswick.

Oostendorp, M. v. (1995). *Vowel quality and phonological projection*. PhD thesis, Tilburg University.

Prince, A. and Smolensky, P. (1993). Optimality theory: Constraint interaction in generative grammar.

Sherer, T. D. (1994). *Prosodic phonotactics*. PhD thesis, University of Massachusetts Amherst.

Smolensky, P. (1996). On the comprehension/production dilemma in child language. *Linguistic inquiry*, 27(4):720–731.

Tesar, B. (2004). Using inconsistency detection to overcome structural ambiguity. *Linguistic Inquiry*, 35(2):219–253.

Tesar, B. (2008). Output-driven maps. Ms, Rutgers University. Available as ROA-956 from the Rutgers Optimality Archive.

Tesar, B. (2012). Learning phonological grammars for output-driven maps. In *The Proceedings of NELS 39*. Cornell University.

Tesar, B. (2013). *Output-driven phonology: Theory and learning*. Cambridge University Press.

Tesar, B. and Smolensky, P. (1998). Learnability in optimality theory. *Linguistic Inquiry*, 29(2):229–268.

Tesar, B. B. (1995). *Computational optimality theory*. PhD thesis, University of Colorado.