

# Similarity Avoidance in the Proto-Indo-European Root

Adam I. Cooper\*

## 1 Introduction

In their 2004 *Natural Language & Linguistic Theory* paper “Similarity avoidance and the OCP,” Frisch, Pierrehumbert, and Broe advance an alternative account of co-occurrence restrictions in the Arabic trilateral root system. The approach relies on a concept of *similarity avoidance* (SA), a “processing constraint that disfavors repetition” (180). Their claim is “lexical items that avoid repetition will be easier to process, and so will be favored in acquisition, lexical borrowing, coining novel forms, and in active usage” (221). For the Arabic data, “all of the co-occurrence restrictions can be captured by a single constraint, similarity avoidance, where the degree of co-occurrence restriction depends on the similarity between homorganic consonants. The strength of the constraint is predicted by the similarity of the consonants that are involved...” (197).

The SA approach, developed basically in two parts—the identification of co-occurrence restrictions through quantification of over- and under-representation in the data, and the appeal to calculated values of similarity as a means of accounting for co-occurrence phenomena—presents an appealing methodology for the study of languages other than, and unrelated to, Arabic. In this paper I seek to probe the explanatory powers of SA for Proto-Indo-European (PIE), the reconstructed parent of the Indo-European language family, in particular the structure of its verbal root system. I begin in section 2 by providing relevant background information on PIE phonology and morphology. In section 3 I formulate the two components of a SA account of the PIE verbal root, and evaluate the results. In section 4 I explore in some detail a particular co-occurrence phenomenon, for which SA ostensibly fails as a means of account. I conclude in section 5, and discuss potential future work on this topic.

## 2 Relevant Aspects of PIE Phonology and Morphology

Important for the current undertaking will be a clear understanding of the consonantal inventory of PIE, as well as basic information concerning the root and restrictions on its shape.

The PIE consonantal inventory (adapted from Mayrhofer, 1986; Meier-Brügger, 2002; Weiss, forthcoming, et al.) is given in (1). For now it will suffice to maintain a certain neutrality in the presentation; details of the necessary stipulations made for the purposes of calculating similarity values are given below in section 3.2, where the feature matrix for PIE consonants is introduced.

(1) The Consonants of PIE

	LABIAL	CORONAL	DORSAL			“LARYNGEAL”
			Palatal	Velar	Labiovelar	
<i>Stops</i>	*p *b *b <sup>h</sup>	*t *d *d <sup>h</sup>	*ʔ *ǵ *ǵ <sup>h</sup>	*k *g *g <sup>h</sup>	*kʷ *gʷ *gʷ <sup>h</sup>	*h <sub>1</sub> *h <sub>2</sub> *h <sub>3</sub>
<i>Fricatives</i>		*s				
<i>Nasals</i>	*m	*n				
<i>Liquids</i>		*l *r				
<i>Glides</i>	*w		*j		*y	

The PIE root is at its core a sequence CVC, with consonant clusters potentially occurring prevocally and/or postvocally. Examination of *Lexikon der Indogermanischen Verben* (henceforth *LIV*; Rix et al. 2001) reveals the following possible shapes for verbal roots:

\*Many thanks to Abby Cohn, Alan Nussbaum, Michael Wagner, Michael Weiss and Draga Zec for very helpful comments and suggestions.

## (2) PIE Root Shapes

Where *e* = fundamental, ablauting vowel and C = a consonant.

- |    |         |                       |         |     |          |                                      |                     |
|----|---------|-----------------------|---------|-----|----------|--------------------------------------|---------------------|
| a. | *CeC-   | *h <sub>1</sub> es-   | ‘be’    | f.  | *CCCeC-  | *pster-                              | ‘sneeze’            |
| b. | *CeCC-  | *tens-                | ‘pull’  | g.  | *CCCeCC- | *streig-                             | ‘penetrate’         |
| c. | *CCeC-  | *g <sup>h</sup> uēn-  | ‘ring’  | (h. | *CCeCCC- | *sp <sup>h</sup> erh <sub>2</sub> g- | ‘crackle’)          |
| d. | *CCeCC- | *h <sub>2</sub> melĝ- | ‘milk’  | (i. | *CCCeCC- | *sptieūH-                            | ‘spit’ <sup>1</sup> |
| e. | *CeCCC- | *menth <sub>2</sub> - | ‘whisk’ |     |          |                                      |                     |

That the PIE root is ostensibly syllabic in shape notably distinguishes it from the Arabic root, a fact which will receive further attention below in section 3.1.

Finally, constraints<sup>2</sup> posited by scholars on the shape of the PIE root (Benveniste, 1935; Kuryłowicz, 1935; Meillet, 1964; Szemerényi, 1996; Weiss, forthcoming, et al.) are given in (3).

(3) PIE Root Consonantal Co-Occurrence Restrictions<sup>3</sup>

- |    |                                      |   |
|----|--------------------------------------|---|
| a. | *C <sub>i</sub> VC <sub>i</sub>      | in which the consonants are identical. <sup>4</sup>   |
| b. | *mVP/PVm                             | in which <i>m</i> is the labial nasal and <i>P</i> is any labial oral stop. <sup>5</sup>                                |
| c. | *DVD                                 | in which <i>D</i> represents a voiced unaspirated stop. <sup>6</sup>  |
| d. | *TVD <sup>h</sup> /D <sup>h</sup> VT | in which <i>T</i> represents a voiceless stop and <i>D<sup>h</sup></i> represents a voiced aspirated stop. <sup>7</sup> |
| e. | *CVRR                                | in which <i>R</i> represents any sonorant consonant.  |

Being traditionally held in the literature, and, more importantly, militating against co-occurrence of segments ostensibly similar in nature—in terms of manner and/or place of articulation—these constraints will serve as a suitable basis of comparison for the results obtained in this study.

### 3 Similarity Avoidance and PIE

#### 3.1 Identifying Co-Occurrence Restrictions in the PIE Root

The identification of co-occurrence restrictions through a more statistically-oriented methodology will seek to evaluate the picture painted by constraints of the sort presented above in (3).<sup>8</sup> Unlike Frisch et al., I will be concerned not only with homorganic consonantal co-occurrence, but with co-occurrence of consonants sharing manner features as well, as this broader scope is better aligned with the nature of the root structure constraints which have been posited for PIE.

The source of the data used for the analysis is the aforementioned *LIV* (Rix et al., 2001), a

<sup>1</sup>The last two shapes are included parenthetically as they are found only once in *LIV*, in the ambiguous forms provided here: \*sp<sup>h</sup>erh<sub>2</sub>g-, which has a voiceless aspirated stop, a rather marginal sound not considered to be phonemic, and \*sptieūH-, which is very likely onomatopoeic, and has a laryngeal of unclear nature.

<sup>2</sup>The term is not necessarily to be taken in an Optimality-Theoretic sense.

<sup>3</sup>Note that one asterisk marks ungrammaticality, the other a reconstructed form.

<sup>4</sup>Hopper (1973) extends the restriction to homorganic consonants in general, stating “the two consonants must differ in point of articulation regardless of the manner feature” (158).

<sup>5</sup>Ringe (1998, 2006) posits a more general constraint such that “a root cannot begin and end with stops of the same place of articulation” (1998, 175) or “a root could not contain oral stops at the same place of articulation both in its onset and in its coda” (2006, 7). To be explicit, perhaps a better characterization for Ringe’s observation might be “a root could not contain [oral] stops at the same place of articulation in any prevocalic or postvocalic position.”

<sup>6</sup>Characterized by Ringe (2006) as “a root could not contain two voiced stops” (8). Note that it in fact could, of course, as long as at least one stop was an aspirate.

<sup>7</sup>Ringe (2006) notes that this sequence could occur, if the voiceless stop followed root-initial \*s (8).

<sup>8</sup>Of course a methodological issue presents itself here, as it is not uncommon practice for posited constraints to in turn be used by scholars in the reconstruction process, and as such, exert influence on the results in a rather more direct, conscious way. The significance of this possibility remains to be fully considered, although it ought not pose a serious threat to the current undertaking.

corpus of 1195 reconstructed PIE verbal roots. Of these 1195 verbal roots, I have chosen to focus on a smaller set of 630, which I take to be reconstructed for PIE with confidence; *LIV* treats the remaining 565 roots as problematic in some way, each being either of questionable PIE date, ambiguous in its reconstruction, or both. Using a corpus of this limited size may ostensibly be cause for concern, but the alternative, incorporating into the analysis forms that lack confidence in their PIE status or reconstruction, is arguably less tenable, and as such will not be pursued here.

The 630 roots in the data set can be divided according to shape as follows: 247 CVCC-, 173 CVC-, 112 CCVC-, 82 CCVCC-, 6 CVCCC-, 5 CCCVCC-, 4 CCCVC- and 1 CCVCCC-. Note in this regard that as per Cser (2001), but contra Matasović (1997 [1999]), postvocalic glides are analyzed as part of the syllable coda, not the nucleus.<sup>9</sup>

I will focus in particular on the co-occurrence of vowel-adjacent consonants, i.e., -CVC-. This approach makes sense for a number of reasons: it allows for the assessment of the majority of the constraints posited for the PIE root in (3), which target the sequence -CVC-; and it maximizes the usefulness of the data set, since all 630 roots contain this sequence as well. Furthermore, it minimizes the clear influence of principles of syllable well-formedness in the phonotactics of the root, which is typically (and understandably) identified as minimally being a syllable of shape CVC (Golston, 1996 in particular). Such a characterization is supported by the simple observation that consonantal sequences pattern in a way which abides by sonority sequencing: in the data set, assuming the sonority hierarchy glides > liquids > nasals > fricatives<sup>10</sup> > stops, the majority of prevocalic biconsonantal sequences rise in sonority (80.36% in CCVC- roots, 78.05% in CCVCC- roots) while the majority of postvocalic biconsonantal sequences fall in sonority (91.50% in CVCC- roots, 89.02% in CCVCC- roots, and 100% in CCCVCC- roots). Focusing exclusively on -CVC- allows for an unobscured study of root co-occurrence, as any restrictions identified will not so readily be explained by sonority sequencing.

The formula we will use is the ratio O(berved)/E(xpected) (Pierrehumbert 1993), given in (4). Following Frisch et al., an O/E less than 1 suggests the existence of a co-occurrence restriction on the consonants concerned; an O/E greater than 1 suggests no such constraint is active.

$$(4) \text{ O/E} = \frac{\text{Observed } \{C_1, C_2\} \text{ co-occurrence in roots}}{\frac{\text{Observed } /C_1/ \text{ occurrence in roots} \times \text{Observed } /C_2/ \text{ occurrence in roots}}{\text{Total roots}}}$$

The frequency of each attested consonant pair in the environment -CVC- in the data set was compared with the frequency of each of its members independently in prevocalic and postvocalic position (not in all positions in the root). The results in Table 1 have been aggregated according to both place and manner of articulation, with deeper distinctions made with respect to the latter than simply sonorant versus obstruent. This is a deviation from the aggregating techniques used by Frisch et al. for their Arabic data (186), but one which is arguably justified given the nature of the PIE data and the shape of the traditional constraints on consonantal co-occurrence in the root.<sup>11</sup>

Looking at these results,<sup>12</sup> we see that pairings of consonants alike in place or manner of articulation are disfavored; in each row of O/E values in either table the lowest O/E value is consistently that in which the prevocalic and postvocalic segments share place or manner, respectively.<sup>13</sup> Focusing first on homorganicity, we observe that it is nearly always the case that the O/E value for like place pairings is below 1, with labial and laryngeal pairings having the lowest values; all unlike place pairings, save for labial-laryngeal, are over 1 (and indeed, labial-laryngeal, with O/E of

<sup>9</sup>If glides were instead part of the nucleus, the CVC- roots would predominate, which would perhaps be more in line with the expectations of the literature, given this shape's canonical status.

<sup>10</sup>Including the three laryngeals; see section 3.2 for explicit identifications assumed for this paper.

<sup>11</sup>I take Frisch et al. to be amenable to an adjustment of this sort for study of languages other than Arabic; for them it is simply the case that "the functional pressures that lead to similarity avoidance in Arabic are released along the place dimension only, and not in manner or voicing" (198, fn. 5).

<sup>12</sup>Statistical significance remains to be fully confirmed, yet the results are nonetheless still provocative.

<sup>13</sup>O/E values lower than 1 occur in the rest of the tables, it should be pointed out, suggesting the existence of co-occurrence restrictions likely not motivated by SA; other forces are presumably at work.

0.82, is close to 1 and higher than any homorganic O/E value). Following Frisch et al., we posit co-occurrence restrictions basically only on homorganic pairings (given  $O/E < 1$ ), and predict that heterorganic pairings occur free of any constraints (given  $O/E > 1$ ). Among the homorganic restrictions, though, there is variation, with the restriction against labial co-occurrence being the strongest ( $O/E = 0.28$ ), and that against dorsals being the weakest ( $O/E = 0.75$ ). As such, consonantal co-occurrence and the restriction thereof must be considered gradient, not categorical.

PLACE OF ARTICULATION		Second Segment (-VC-)			
		Labials	Coronals	Dorsals	“Laryngeals”
First Segment (-CV-)	Labials	<b>0.28</b>	1.42	1.00	0.82
	Coronals	1.45	<b>0.69</b>	1.06	1.15
	Dorsals	1.04	1.01	<b>0.75</b>	1.32
	“Laryngeals”	1.11	1.11	1.13	<b>0.33</b>

  

MANNER OF ARTICULATION		Second Segment (-VC-)				
		Stops	Fricatives	Nasals	Liquids	Glides
First Segment (-CV-)	Stops	<b>0.79</b>	0.82	1.31	1.35	0.80
	Fricatives	1.23	<b>0.50</b>	1.00	1.21	1.07
	Nasals	0.98	1.20	<b>0.67</b>	0.98	1.05
	Liquids	1.05	1.39	0.89	<b>0.00</b>	1.63
	Glides	1.19	1.21	0.68	1.1	<b>0.60</b>

Table 1. -CVC- consonantal pairings, aggregated by place and manner of articulation.

The contrast in degree of restriction across consonantal pairings of like and unlike manner of articulation is overall less robust. This is not only because O/E values for like manner pairings are generally higher than those for like place pairings, but also because there are more instances of O/E values less than 1 occurring throughout the range of possible consonantal pairings (although, the majority of these are relatively close to 1). As such the result here too is a more gradient distribution of co-occurrence restrictions, following Frisch et al.’s guidelines, in which such restrictions apply not only to like pairings of consonants, but unlike pairings as well.

This is not to say that the constraints on root shape traditionally posited for PIE are not without some degree of justification. Looking at O/E values relevant for these constraints, as in Table 2, we see that each constraint is usually matched by a rather low O/E.

Constraint	O/E Value (CVC-, 173)	O/E Value (non-CVC-, 457)	O/E Value (All, 630)	Exceptions <sup>14</sup>
a. * *C <sub>i</sub> VC <sub>i</sub>	<b>0.01</b>	<b>0.01</b>	<b>0.01</b>	* <i>ses-</i> ‘sleep’
b. * *mVP	0.00	0.00	0.00	* <i>skek-</i> ‘move fast’
c. * *PVm	0.00	0.00	0.00	* <i>tetk-</i> ‘produce’
d. * *DVD	0.00	0.00	0.00	* <i>h<sub>1</sub>eh<sub>1</sub>s-</i> ‘sit’
e. * *TVD <sup>h</sup>	0.00	<b>0.92</b>	<b>0.37</b>	* <i>steb<sup>h</sup>-</i> ‘solidify’
f. * *D <sup>h</sup> VT	0.00	0.00	0.00	* <i>skab<sup>h</sup>-</i> ‘scrape’

Table 2. Traditional constraints on the PIE root with three relevant O/E values.

While the adequacy of these constraints as a means of capturing co-occurrence phenomena in the PIE root might now be questionable, given the fact that co-occurrence is generally more gradient than these categorical statements convey, we can at least appreciate their validity in capturing

<sup>14</sup>Some notes on these exceptions: \**ses-* is said to be onomatopoeic, and \**tetk-* and \**h<sub>1</sub>eh<sub>1</sub>s-* are likely products of reduplication; as for \**skek-*, its contiguous distribution only in Germanic, Slavic and Celtic may make its PIE status questionable, although it is not so annotated in *LIV*. Concerning \**steb<sup>h</sup>-* and \**skab<sup>h</sup>-*, see footnote 9.

some of the most restricted pairings. While there is gradience in the system overall, the pairings targeted by these constraints are just about as categorically constrained as possible, given the O/E values at or very near to 0 (the two regular exceptions to  $**TVD^h$  notwithstanding).

In this vein, it is worth mentioning an additional pairing of consonants which is also very strongly restricted: the two PIE liquids  $*l$ ,  $*r$ . The O/E for the co-occurrence of these two sounds is 0: none of the 630 roots under examination, nor any in *LIV* at all, features these two consonants co-occurring, not only in the particular environment -CVC-, but in any combination of pre- and postvocalic positions, immediately adjacent to the root vowel or not. Given the relatively high frequency of each of these consonants in the data— $*r$  occurs 74 times immediately prevocally, 95 times immediately postvocally;  $*l$  occurs 59 and 51 times in these two positions, respectively—this finding is rather striking. A constraint targeting the liquids would, under the traditional approach, seem to warrant introduction; for whatever reason, though, the restrictedness of liquid co-occurrence has generally gone unacknowledged in the literature, save for references such as that by Ringe (1998, 175).

Moving on, we can also appreciate that the pairings above are singular, in the sense that pairings of analogous consonants differing in place of articulation or in some aspect of manner of articulation are typically not so strongly disfavored. The O/E values in Table 3 confirm as much:

Shape	O/E Value (CVC-, 173)	O/E Value (non-CVC-, 457)	O/E Value (All, 630)
a. $**nVT$	<b>0.00</b>	1.69	<b>0.71</b>
b. $**TVn$	<b>0.96</b>	1.28	1.26
c. $**TVT$	1.60	<b>0.98</b>	1.31
d. $**D^hVD^h$	2.06	2.31	2.46
e. $**TVD$	<b>0.40</b>	<b>0.00</b>	<b>0.34</b>
f. $**DVT$	<b>0.58</b>	0.51	<b>0.53</b>
g. $**DVD^h$	<b>0.00</b>	<b>0.00</b>	<b>0.00</b>
h. $**D^hVD$	1.72	<b>0.00</b>	1.91

Table 3. O/E comparisons for analogous -CVC- shapes.

Here we see that not every pairing is necessarily fully permitted ( $O/E > 1$ ), but rather, there is variation in degree of acceptability, both within and across roots of differing shape. For instance, the co-occurrence of coronal oral and nasal stops (a.-b.; note the special use here of cover symbol *T*), is noticeably less constrained than that of their labial counterparts; similar findings hold for voiceless unaspirated stops (c.) and voiced aspirated stops (d.). Combinations of voiceless and voiced unaspirated stops (e.–f.), and voiced unaspirated and aspirated stops (g.–h.), however, are generally disfavored ( $O/E < 1$ ), though notably not to the extent that the co-occurrence of voiced unaspirated stops is.

The traditional account of consonantal co-occurrence in the PIE root has relied on individual constraint statements which isolate those pairings of consonants apparently most noticeably absent in the corpus. While I have shown that there is some legitimacy in the positing of categorical constraints targeting these particular pairings, I have also sought to emphasize that this approach fails to take into account or treat the behavior of other pairings, which are either as categorically impermissible (as in the case of the liquids), or, more commonly, less rigorously restricted, but nonetheless not fully permissible either (as in the case of consonants sharing either place or manner of articulation). In any case, we will now seek to evaluate the extent to which calculated values of similarity can be relied upon as means of capturing this situation.

### 3.2 Employing Similarity Avoidance for PIE

The formula we will use to calculate similarity is Frisch et al.'s “natural classes similarity metric”:

$$(5) \quad \text{Similarity} = \frac{\text{Shared natural classes}}{\text{Shared natural classes} + \text{Non-shared natural classes}} \quad (\text{their (7); 198})$$

A methodological distinction will be made for the study of PIE, in which consonantal co-occurrence is constrained not only by shared place, but by shared manner of articulation as well, unlike Arabic. The calculated similarity values reflect this fact: two heterorganic consonants will not automatically be non-similar (similarity = 0), as they were for Arabic, but will indeed be similar to the extent dictated by their common membership in non-place-based natural classes.

The feature matrix for the PIE consonant inventory is given below in Figure 1. A minimal number of features have been introduced to capture all the distinctions holding between the consonants of PIE. Given the need for specificity in detail, explicit choices have been made about the nature and identity—and by extension, representation in the feature matrix—of some consonants, for which there is a notable degree of uncertainty in reconstruction. The stipulations made here ought not to be considered definitive, by any means—indeed, one of the virtues of the methodology is its potential to accommodate, with only modest effort, other possibilities as well.

	<i>p</i>	<i>b</i>	<i>bʰ</i>	<i>m</i>	<i>w</i>	<i>t</i>	<i>d</i>	<i>dʰ</i>	<i>s</i>	<i>n</i>	<i>l</i>	<i>r</i>	<i>k</i>	<i>g</i>	<i>gʰ</i>	<i>kʷ</i>	<i>gʷ</i>	<i>gʷʰ</i>	<i>j</i>	<i>h</i>	<i>ʕ</i>	<i>h</i>		
cons	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	
son	-	-	-	+	+	-	-	-	-	+	+	+	-	-	-	-	-	-	-	-	+	+	+	+
cont	-	-	-	-	+	-	-	-	+	-	+	+	-	-	-	-	-	-	-	-	+	+	+	+
nas	-	-	-	+		-	-	-		+			-	-	-	-	-	-	-	-				
lat											+	-												
lab	+	+	+	+	+																			
cor						+	+	+	+	+	+	+												
dors													+	+	+	+	+	+	+	+				
phar																					+	+	+	
rad																					+	+		
voi	-	+	+	+	+	-	+	+	-	+	+	+	-	+	+	-	+	+	+	+	-	+		
mid					+								-	-	-	-	-	+	+	+				
high													-	-	-	+	+	+	-	-	-			
sp	-	-	+			-	-	+					-	-	+	-	-	+	-	-				
glot																							+	

Figure 1. Feature matrix for PIE consonants.

Two sets of featural representations deserve closer attention. First, regarding the three dorsal stop series, I assume plain velars to be simply velars, palatovelars  $*k, g, g^h$  to be velar stops with secondary palatal articulation (“palatalized velars”  $/k^j, g^j, g^{jh}/$ ), and labiovelars  $*k^w, g^w, g^{wh}$  to be, similarly, velar stops with secondary labial articulation (“labialized velars”  $/k^w, g^w, g^{wh}/$ ).<sup>15</sup> Following Lahiri and Evers (1991), palatovelars are characterized as [+high] and labiovelars as [+round], and plain velars as [-high, -round]. As for the PIE “laryngeals”  $*h_1, h_2, h_3$ , I simply and explicitly assume the following identifications, and associated feature matrices, in accordance with some of the speculative statements of Fortson (2004) and Weiss (forthcoming): the first laryngeal as the voiceless glottal fricative  $/h/$ , the second laryngeal as the voiceless pharyngeal fricative  $/ħ/$ , and the third laryngeal as the voiced pharyngeal fricative  $/ʕ/$ .

Based on the 82 unique natural classes assembled using this feature matrix, similarity values were calculated for all the consonants of the language; they are presented in Table 4. We see that similarity is higher for homorganic consonants than for consonants of like manner. This correlates nicely with the calculated O/E values, which showed restrictions on shared place to be stronger than those on shared manner. Further, if we focus on the constraints categorically posited for PIE—which may remain relevant because, again, they involve pairings of consonants most constrained in the root—we find that SA, as developed here, can indeed account for most of them. The constraint on identical consonants is easily explained, because the relevant consonants share the highest degree of similarity possible, 1—absolute identity. Also, the relatively higher similarity values holding within the class of labial non-continuants, due to its smaller size as compared to,

<sup>15</sup>The phonetic identification of the labiovelars is to my understanding less tenuous than that of the palatovelars, which have also been identified with, for instance, plain palatal stops. It is also worth noting that the identification of the “plain velars” as the basic series is not itself an uncontroversial determination.

for instance, the coronals, account for the greater degree of restricted co-occurrence between its members. (Both of these results are similar to those obtained by Frisch et al. for Arabic.) Finally, the constraint on liquid co-occurrence is accounted for as well, given the strength of the similarity of *\*l* and *\*r* (0.83), which differ only in the single feature [ $\pm$ lateral].

	<i>p</i>	<i>b</i>	<i>b<sup>h</sup></i>	<i>m</i>	<i>w</i>	<i>t</i>	<i>d</i>	<i>d<sup>h</sup></i>	<i>s</i>	<i>n</i>	<i>l</i>	<i>r</i>
<i>p</i>	1.00	0.45	0.33	0.19	0.05	0.37	0.20	0.16	0.17	0.09	0.05	0.05
<i>b</i>		1.00	0.65	0.36	0.12	0.19	0.35	0.26	0.08	0.15	0.08	0.08
<i>b<sup>h</sup></i>			1.00	0.38	0.13	0.15	0.26	0.38	0.09	0.16	0.08	0.08
<i>m</i>				1.00	0.33	0.08	0.15	0.15	0.05	0.37	0.20	0.20
<i>w</i>					1.00	0.00	0.04	0.04	0.05	0.14	0.22	0.22
<i>t</i>						1.00	0.45	0.35	0.35	0.17	0.09	0.09
<i>d</i>							1.00	0.67	0.17	0.35	0.16	0.16
<i>d<sup>h</sup></i>								1.00	0.18	0.36	0.17	0.17
<i>s</i>									1.00	0.10	0.25	0.25
<i>n</i>										1.00	0.41	0.41
<i>l</i>											1.00	0.83
<i>r</i>												1.00

	<i>k</i>	<i>g</i>	<i>g<sup>h</sup></i>	<i>k<sup>j</sup></i>	<i>g<sup>j</sup></i>	<i>g<sup>jh</sup></i>	<i>k<sup>w</sup></i>	<i>g<sup>w</sup></i>	<i>g<sup>wh</sup></i>	<i>j</i>	<i>h</i>	<i>ʕ</i>	<i>h</i>
<i>p</i>	0.37	0.20	0.17	0.37	0.20	0.17	0.35	0.19	0.16	0.00	0.00	0.06	0.00
<i>b</i>	0.19	0.35	0.27	0.19	0.35	0.27	0.19	0.33	0.26	0.04	0.00	0.00	0.04
<i>b<sup>h</sup></i>	0.15	0.26	0.39	0.15	0.26	0.39	0.15	0.25	0.38	0.04	0.05	0.00	0.05
<i>m</i>	0.08	0.15	0.16	0.08	0.15	0.16	0.08	0.14	0.15	0.17	0.06	0.06	0.18
<i>w</i>	0.00	0.04	0.04	0.00	0.04	0.04	0.04	0.07	0.08	0.46	0.21	0.21	0.38
<i>t</i>	0.33	0.19	0.15	0.33	0.19	0.15	0.32	0.18	0.15	0.00	0.00	0.05	0.00
<i>d</i>	0.19	0.33	0.26	0.19	0.33	0.26	0.18	0.32	0.25	0.04	0.00	0.00	0.04
<i>d<sup>h</sup></i>	0.15	0.25	0.38	0.15	0.25	0.38	0.14	0.24	0.36	0.04	0.05	0.00	0.04
<i>s</i>	0.15	0.08	0.09	0.15	0.08	0.09	0.14	0.08	0.08	0.06	0.07	0.15	0.07
<i>n</i>	0.08	0.15	0.16	0.08	0.15	0.16	0.08	0.14	0.15	0.17	0.06	0.06	0.18
<i>l</i>	0.04	0.07	0.08	0.04	0.07	0.08	0.04	0.07	0.08	0.27	0.13	0.13	0.29
<i>r</i>	0.04	0.07	0.08	0.04	0.07	0.08	0.04	0.07	0.08	0.27	0.13	0.13	0.29
<i>k</i>	1.00	0.45	0.30	0.65	0.33	0.25	0.61	0.32	0.24	0.05	0.00	0.05	0.00
<i>g</i>		1.00	0.55	0.33	0.64	0.42	0.32	0.61	0.40	0.08	0.00	0.00	0.04
<i>g<sup>h</sup></i>			1.00	0.25	0.42	0.68	0.24	0.40	0.65	0.09	0.05	0.00	0.05
<i>k<sup>j</sup></i>				1.00	0.45	0.30	0.61	0.32	0.24	0.05	0.00	0.05	0.00
<i>g<sup>j</sup></i>					1.00	0.55	0.32	0.61	0.40	0.08	0.00	0.00	0.04
<i>g<sup>jh</sup></i>						1.00	0.24	0.40	0.65	0.09	0.05	0.00	0.05
<i>k<sup>w</sup></i>							1.00	0.48	0.33	0.05	0.00	0.05	0.00
<i>g<sup>w</sup></i>								1.00	0.57	0.08	0.00	0.00	0.04
<i>g<sup>wh</sup></i>									1.00	0.09	0.05	0.00	0.04
<i>j</i>										1.00	0.27	0.27	0.50
<i>h</i>											1.00	0.5	0.44
<i>ʕ</i>												1.00	0.44
<i>h</i>													1.00

Table 4. Similarity of PIE consonants, organized by place of articulation.

SA is unable to account for all instances of restricted consonantal co-occurrence, though. The constraint  $*\text{TVD}^h$  understandably defies explanation by appeal to SA, as voiceless unaspirated stops and voiced aspirated stops are the least similar to each other of the three PIE stop series; each are more similar to their voiced unaspirated counterparts. As such, it can hardly be the case that similarity conditions their lack of co-occurrence. It seems, then, that other forces interact with the drive to avoid similarity in the PIE system, to yield this result. On the other hand, the posited PIE root constraint  $*\text{DVD}$  would appear to be entirely amenable to a SA analysis, as it restricts the co-occurrence of members of a single natural class of consonants, voiced unaspirated stops. Surprisingly, however, this does not seem to be the case, a finding which is explored in greater detail in the next section.

#### 4 The Constraint $*\text{DVD}$ : A Special Problem

The only PIE stop series to have a co-occurrence constraint posited in the literature is the voiced unaspirated stops. This position is supported by consideration of relevant O/E values: voiced unaspirated stops do not co-occur at all in the root (O/E = 0), whereas voiceless unaspirated stops and voiced aspirated stops co-occur freely (O/E = 1.31, 2.46 respectively).

If SA is to account for this asymmetry, we expect the similarity values for voiced unaspirated stops to be markedly higher than those for the two other stop series; co-occurrence of voiced unaspirated stops could understandably be impermissible as a result. This is not the case, however, as shown in (6):

##### (6) Similarity Values for the Three PIE Stop Series

a. Voiceless unaspirated (O/E = 1.33)						b. Voiced unaspirated (O/E = 0.00)					
	$*p$	$*t$	$*k$	$*k^h$		$*b$	$*d$	$*g$	$*g^h$	$*g^h$	
$*p$	1.00	0.37	0.37	0.37	0.35	$*b$	1.00	0.35	0.35	0.35	0.33
$*t$		1.00	0.33	0.33	0.32	$*d$		1.00	0.33	0.33	0.32
$*k$			1.00	0.65	0.61	$*g$			1.00	0.64	0.61
$*k^h$				1.00	0.61	$*g^h$				1.00	0.61
					1.00	$*g^h$					1.00

  

c. Voiced aspirated (O/E = 2.46)					
	$*b^h$	$*d^h$	$*g^h$	$*g^h$	$*g^h$
$*b^h$	1.00	0.38	0.39	0.39	0.38
$*d^h$		1.00	0.38	0.38	0.36
$*g^h$			1.00	0.68	0.65
$*g^h$				1.00	0.65
$*g^h$					1.00

All three stop series have similar similarity values holding among their members, and in fact, the values for the voiced unaspirated series are generally the lowest. How can SA be invoked as an explanation, then? What differentiates voiced unaspirated stops from the other two series, such that similar constraints do not hold for them as well?

Indeed, issues like this one concerning the constraint  $*\text{DVD}$  have led some to question its validity as a true restriction on root shape in the grammar of PIE. Iverson and Salmons (1992), and later Barrack (2002), for example, have attempted to explain away the effects of  $*\text{DVD}$  as a result of distributional phenomena manifested by stops in general and voiced unaspirated stops in particular. For Iverson and Salmons (301–304), these are that 1) PIE double-stop roots are of “striking rarity” (301); 2) voiced stops in general are less attested in PIE; and 3) PIE  $*b$  is mark-



edly rare. Barrack (82–84) reiterates the last two points and adds another: 4) stops occur less frequently in the coda of a syllable than in the onset.

I would like to consider this idea further by pointing out a few facts about consonantal co-occurrence in the PIE root in general which will place it in what I think is much-needed context. As points 3) and 4) are confirmed by examination of the data set used here, I will not comment on them further; I will focus on points 1) and 2), beginning with the latter.

Voiced unaspirated stops in general are about as well attested in PIE as their aspirated counterparts. In the 630 roots in the data set, voiced unaspirated stops occur a total of 150 times, while voiced aspirated stops occur a total of 167 times. If we break down the counts by place of articulation, as in Table 5, we see that what is behind their near equivalency (at least as compared to the voiceless stops, which number 318) is the scarcity of *\*b*; if it occurred with a frequency more like that of the other voiced unaspirated stops, the numbers would clearly shift.

LABIAL		CORONAL		VELAR		PALATOVELAR		LABIOVELAR	
<i>*p</i>	80	<i>*t</i>	85	<i>*k</i>	65	<i>*kʰ</i>	31	<i>*kʷ</i>	57
<i>*b</i>	4	<i>*d</i>	69	<i>*g</i>	27	<i>*gʰ</i>	21	<i>*gʷ</i>	29
<i>*bʰ</i>	53	<i>*dʰ</i>	52	<i>*gʰ</i>	23	<i>*gʷʰ</i>	13	<i>*gʷʰ</i>	26

Table 5. The frequency of stops in all positions in the data.

To identify the relative rarity of voiced unaspirated stops, then, as a factor in the rarity of roots of shape *\*DVD* is not so simple; something ought to be said about why voiced aspirated stops do not behave the same way and are similarly restricted.

Secondly, with respect to the “striking rarity” of double-stop roots in PIE, this claim isolates a single aspect of a larger picture, namely, the general rarity of roots with two consonants of like manner in pre- and postvocalic positions. If we consider again the O/E values in Table 1 above, we see that co-occurrence in terms of like manner of articulation is restricted across the board. Indeed, co-occurrence of stops appears to be least restricted; if we subdivide the results along the lines of the three stop series (the relevant O/Es are included above in (6)), we see that it is the non-restricted co-occurrence of voiceless unaspirated stops and voiced aspirated stops (i.e., *\*TVT* and *\*DʰVDʰ*) which yields this result. Within the general manner classes, it is only these two stop series which constitute an aberration, then; the restricted co-occurrence of voiced unaspirated stops is actually, in a sense, typical. Given these findings, it is not really voiced unaspirated stops whose co-occurrence deserves special attention, but rather it is voiceless unaspirated stops and voiced aspirated stops whose idiosyncratic behavior deserves closer scrutiny.

I leave it to future work to explore the issue further from this perspective. For our current purposes, it is enough to say that the inability of the SA analysis as developed here to account for the posited constraint *\*DVD* ought not to be taken as a detraction outright, but rather, the issue lies more with whether this constraint ought to be posited in the first place.

## 5 Conclusion

In this paper I have extended the similarity avoidance approach of Frisch, Pierrehumbert, and Broe (2004) to the study of the Proto-Indo-European verbal root system, with some careful adjustment in the methodology. The picture which emerges of consonantal co-occurrence in the PIE root, in examining calculated O(berved)/E(xpected) values, is much more fine-grained than traditional root structure constraints alone can convey. Restrictions hold throughout the system, and both place and manner of articulation play a role; the nexus for their influence on non-identical consonants is best seen in the absolute lack of co-occurrence of the liquids.

Calculated values of similarity can generally be invoked to account for the observed co-occurrence phenomena: homorganic consonants are more similar than consonants of like-manner, which correlates with the O/E findings. Still, the inability of straightforward SA to account for the constraint on voiced unaspirated stops is surprising, given the ostensible influence of similarity on it, and evokes the claim of Iverson and Salmons (1992) and Barrack (2002) about the invalidity of *\*DVD*. That their claim leaves unacknowledged some broader issues concerning consonantal

co-occurrence renders it not entirely satisfying, however; the issue requires further consideration.

Finally, in subjecting PIE to a means of analysis afforded by current approaches in phonological theory—an exercise which has been attempted previously, though not unproblematically, in, for instance, the statistical study of PIE root forms by Jucquois (1966) and work promoting the Glottalic Theory (Hopper, 1973; Gamkrelidze & Ivanov, 1973, et al., but see Barrack, 2002 for refutation)—this study not only sheds new light on the phonotactic restrictions of PIE, but it also seeks to bridge the methodological gap between traditional PIE scholarship and current phonological theory. Expanding the scope of the PIE SA account into the diachronic domain would do more in this regard: applying the methodology used here to the IE daughter languages, and comparing the results with those obtained in the present study, could shed light on the extent to which SA may be a driving force in language change in general.

## References

- Barrack, Charles M. 2002. The Glottalic Theory revisited: a negative appraisal. *Indogermanische Forschungen* 107:76–95.
- Benveniste, Emile. 1935. *Origines de la formation des noms en indo-européen*. Paris: Librairie Adrien-Maisonneuve.
- Cser, András. 2001. Diphthongs in the syllable structure of Latin. *Glotta* 75:172–193.
- Fortson, Benjamin W. 2004. *Indo-European Language and Culture: An Introduction*. Malden, Mass.: Blackwell.
- Frisch, Stefan A., Janet B. Pierrehumbert and Michael B. Broe. 2004. Similarity avoidance and the OCP. *Natural Language & Linguistic Theory* 22:179–228.
- Gamkrelidze, Thomas V., and Vjacheslav V. Ivanov. 1972. Sprachtypologie und die Rekonstruktion der gemeinindogermanischen Verschlüsse. *Phonetica* 27:150–156.
- Golston, Chris. 1996. Prosodic constraints on roots, stems and words. In *Interfaces in Phonology (Studia Grammatica 41)*, ed. U. Kleinhenz, 172–193. Berlin: Akademie Verlag.
- Hopper, Paul J. 1973. Glottalized and murmured occlusives in Indo-European. *Glotta* 7:141–166.
- Iverson, Gregory K., and Joseph C. Salmons. 1992. The phonology of the Proto-Indo-European root structure constraints. *Lingua* 87:293–320.
- Jucquois, Guy. 1966. La structure des racines en indo-européen envisage d'un point de vue statistique. In *Linguistic Research in Belgium*, ed. Y. Lebrun, 57–68. Wetteren: Universa.
- Kurylowicz, Jerzy. 1935. *Etudes indoeuropéennes*. Krakow: Gebethner and Wolff.
- Lahiri, Aditi, and Vincent Evers. 1991. Palatalization and coronality. In *The Special Status of Coronals: Internal and External Evidence (Phonetics and Phonology 2)*, ed. C. Paradis and J.-F. Prunet, 79–100. San Diego: Academic Press.
- Matasović, Ranko. 1997 [1999]. The syllabic structure of Proto-Indo-European. *Suvremena lingvistika* 43–44:169–184.
- Mayrhofer, Manfred. 1986. *Indogermanische Grammatik, Band I: Lautlehre*. Heidelberg: Winter.
- Meier-Brügger, Michael. 2002. *Indogermanische Sprachwissenschaft*. Berlin, New York: Walter de Gruyter.
- Meillet, Antoine. 1964. *Introduction à l'étude comparative des langues indo-européennes*. 8th ed. Tuscaloosa, Ala.: University of Alabama Press.
- Pierrehumbert, Janet. 1993. Dissimilarity in the Arabic verbal roots. *Proceedings of the Annual Meeting of the North Eastern Linguistics Society* 23, 67–381. Amherst, Mass.: GLSA, UMass/Amherst.
- Ringe, Don. 1998. Probabilistic evidence for Indo-Uralic. In, *Nostratic: Sifting the Evidence*, ed. J. C. Salmons and B. D. Joseph, 153–197. Amsterdam, Philadelphia: John Benjamins.
- Ringe, Don. 2006. *A Linguistic History of English, Volume I: From Proto-Indo-European to Proto-Germanic*. Oxford, New York: Oxford University Press.
- Rix, Helmut, et al., ed. 2001. *Lexikon der indogermanischen Verben*. Wiesbaden: L. Reichert.
- Szemerényi, Oswald. 1996. *Introduction to Indo-European Linguistics*. Oxford: Clarendon.
- Weiss, Michael. Forthcoming. *Outline of the Historical and Comparative Grammar of Latin*. Ann Arbor: Beech Stave.

Department of Linguistics  
Cornell University  
Ithaca, NY 14853-4701  
ac244@cornell.edu