Similar Place Avoidance: A Statistical Universal

Konstantin Pozdniakov, Guillaume Segerer

▶ **To cite this version:**

Konstantin Pozdniakov, Guillaume Segerer. Similar Place Avoidance: A Statistical Universal. Linguistic Typology, 2007, 11 (2), pp.307-348. halshs-00255870

HAL Id: halshs-00255870
https://shs.hal.science/halshs-00255870

Submitted on 1 Sep 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Similar Place Avoidance: A Statistical Universal*

Konstantin POZDNIAKOV & Guillaume SEGERER
LLACAN (CNRS, INALCO, Université Paris 7)

To †Sergei Starostin

ABSTRACT

In recent years there has been interest in the phenomenon of "similar place avoidance" (SPA), particularly as concerns Arabic CCC radicals. Although little evidence has been presented outside Arabic, Hebrew, and perhaps Semitic in general, where roots with successive consonants sharing the same place of articulation are underrepresented, Frisch, Pierrehumbert & Broe (2004) hypothesize that similarity avoidance may represent a universal tendency. Originally unaware of the work of Pierrehumbert and her co-workers, we undertook an extensive cross-linguistic investigation of SPA and found impressive support for this notion. Whereas the Frisch et al study is largely interested in demonstrating the synchronic reality of SPA in Arabic, the present study originally derived from the diachronic concerns we have, particularly as concerns the subgrouping and historical relations between the diverse Niger-Congo languages spoken in Sub-Sarahan Africa. As documented in the following discussion, these diachronic concerns led us to move progressively further out from our primary concern, the Atlantic subgroup of Niger-Congo, arriving first at non-Niger-Congo Africa and then eventually outside of Africa itself. The result is a dramatic confirmation of SPA effects as a linguistic universal.

0. Introduction

Based on what we know about phonological systems, it seems reasonable to suppose that the major place features of two consonants which make up a -C_1VC_2- root should be independent of each other. After all, no language has been reported where, say, a C_2 coronal consonant assimilates in labiality to a C_1 labial consonant, or vice-versa. Thus, an input such as /bat/ is never realized *[bap] or *[dat]. Major place assimilation is not expected to apply across a vowel.[1] While opposite dissimilatory processes affecting place of articulation are attested (Grammont 1895), they are rare and seldom regular. We therefore do not expect an input such as /bap/ to be realized *[bat] or *[dap] by regular phonological rule. Given the relative independence of major features on consonants separated by a vowel, it comes as somewhat of a surprise that there are statistical biases in which transvocalic consonants can succeed each other within roots. Specifically, in a number of languages which are discussed below, we have noted that two consonants produced at the same place of articulation are significantly underrepresented in lexical -CVC- sequences. Below we report on statistical studies we have done on more than 30 genetically, typologically, and geographically diverse

* The ideas developed here were first presented at the Workshop on Proto-Niger-Congo held in Paris, October 11-16, 2004, which was organized by the Santa Fe Institute and the CNRS-LLACAN in the context of the Evolution of Human Languages Project directed by †Sergei Starostin. We would like to express our grateful feelings to Larry Hyman. Not only has he translated this paper from French, but his suggestions, comments, and encouragements at every stage of this paper have been especially helpful to us. Any possible errors or misinterpretations remain entirely our responsiblity.

[1] As has been recently documented by Hansson (2001) and Rose & Walker (2004), non-adjacent consonant harmony is typically limited to nasal, laryngeal, and secondary coronal features such as anteriority and retroflection. Major place harmony is of course widely attested in child language (see, for example, Pater & Werle 2001 and references cited therein).

languages. Our calculations reveal a striking regularity in the underrepresentation of homorganic consonants in -CVC- sequences.

Such distributional irregularities involving consonant place have been long noted in Semitic studies, particularly as concerns Arabic, whose triliteral √CCC roots avoid consonants at the same place of articulation (Greenberg 1950, Fleisch 1961). The Arabic instantiation of similar place avoidance (henceforth, SPA) has been studied in great detail by Frisch (1996) and Frisch, Pierrehumbert & Broe (2004), who also demonstrate that speakers are aware of such statistical biases, which they relate to the Obligatory Contour Principle (OCP): "Adjacent identical elements are prohibited" (McCarthy 1986: 208). Some recent detailed statistical studies have also confirmed SPA in Japanese (Kawahara *et al.*, 2005), in Muna (Coetzee & Pater 2006) and in Proto-Bantu (Teil-Dautrey, to appear).

The purpose of the present article is to show that the SPA phenomenon is not a specific property of Arabic, Japanese or other isolated languages, but is in fact observed in most, if not all languages of the world. It should be noted that this result is not the confirmation of an a priori theoretical postulate. We approached this issue with no bias—in fact, we stumbled on it quite by accident. Surprised to discover SPA in a number of languages in the Atlantic sub-branch of Niger-Congo,[2] we believed we were dealing with an inherited genetic trait. In order to verify the Atlantic hypothesis, we felt compelled to investigate possible SPA effects in other languages. In the process, and based on extensive cross-linguistic testing, we arrived at our current position, that (statistical) SPA is a likely universal property of human language.

The remainder of our paper is organized as follows. §1 introduces the circumstances which initially led us to conduct our statistical counts as well as the statistical techniques we adopted. The following three sections (§2, §3, §4) respectively explore Atlantic languages, other Niger-Congo languages, and non-Niger-Congo African languages. In order to show that we are not dealing only with an African areal phenomenon, we document SPA in a few non-African languages in §5. Here we also address the question of how our findings might have been affected in languages which we do not know well, particularly as concerns the morphological structure of the lexical items used in our statistical analyses. In the final discussion (§6) and conclusion (§7) sections, we consider the implications of our finding, presenting hypotheses and raising questions for future research.

1. The problem and methodology

The problem under discussion in this paper attracted our attention during the course of preparing a lexical corpus of a group of West African languages which belong to the Atlantic sub-branch of Niger-Congo. Our purpose was to do lexical comparison for the purpose of subgrouping and ultimate reconstruction. In fact, the Atlantic languages are very heterogeneous, lexically, and although most specialists continue to treat them as belonging to a single Niger-Congo sub-branch, lexicostatistical counts often result in low cognate counts almost at the level of chance (e.g. 5 à 7 % based on the Swadish 100-word list; cf. Sapir 1971). To take one example, an allegedly stable notion such as 'big' produces several dozens of different roots among the 40 Atlantic languages we examined. Confronted by such variation, we developed a procedure which would permit us to display lexical information in such a way as to reveal possible formal relationships

[2] For more information on Niger-Congo and its sub-branches, Williamson & Blench (2000) and the chapters in Bendor-Samuel (1989).

between the lexical items in question: For each notion, we constructed a table where the languages are listed in the first column and the diverse consonant combinations (C_1-C_2) head the additional columns. The cells of the table are filled by the roots themselves, based on the place of articulation of their C_1 and C_2.

As a first step in preparing these tables, we assigned each phoneme of each root to one of the major place classes: P, T, C, K. Thus, the phonemes /p, b, f, v, w, m, mb, ɓ/ are represented by the symbol P, which represents the class of labial consonants. Similarly, dental (and alveolar) consonants are symbolized as T, (alveo-)palatal consonants as C, and velar consonants as K. By this procedure, the consonants of all the examined roots will have been assigned to one of the symbols P, T, C, K. Since the majority of the lexical roots examined in these languages have the structure -CVC-, we therefore have a total of 16 possible combinations based on the four major classes, P, T, C and K. These 16 combinations are the 16 columns which follow the language names in the table. Table 1 illustrates this procedure for the notion 'hair'.

	P-P	**P-T**	**P-C**	**P-K**	**T-P**	**T-T**	**T-C**	**T-K**	**C-P**	**C-T**	**C-C**	**C-K**	**K-P**	**K-T**	**K-C**	**K-K**
Fula		waːr			leɓ										gac	
Sereer		wiil														
Basari		mban, fur														
Bedik		mbal	mboy													
Konyagi		muul														
Pen		mban														
Ndut		fen											xoɓ			
Noon		fen														
Safen		fan														
Lehar		mul														
Palor		fen														
Wolof		*war							*jaaw			cok				
Buy				bunk				dung					kum	gen		
Nyun														gen		
Biafada			wey													
Balanta												yɛɛg		hul		
Joola 1		wal							jab							
Joola 2		wan														
Manjaku		wel, faal														
Mankañ		wel, fal							jab					gaal		
Pepel										yeel						
Bijogo		wen														
Nalu		fɛl			lew											
Nalu					θob											
Sua			wiŋ													
Baga K.		foon														
Baga M.		foon														
Landuma		foon														
Temne		fon														
Bullom								ring								
Kisi										yin						
Sherbro								ding	zem							
Gola					dum											

Table 1 - Words for 'hair' in the Atlantic languages

As can be seen in Table 1, the roots which are grouped together in the same column are not necessarily related. In addition, depending on sound changes, historically related roots may appear in different columns. Despite this, there is a greater probability that related roots will be found in the same rather than a different column. This is the justification of the procedure.

We have applied this method and constructed this type of table for numerous basic lexical notions. In the course of this we have noticed that certain columns, on average, seem "more empty" than others. For example, despite the numerous reflexes seen in Table 1, the column K-K is completely devoid of words for 'hair', but also is striking empty for other basic lexical notions as well. For a given gloss, it would be completely normal if certain columns were empty. The closer the attested reflexes are of an ultimate proto-root, the fewer filled columns there should be. However, it was quite surprising to us that for dozens of basic notions, we find almost the same columns empty.

What does this fact mean? Should one conclude that the Atlantic languages avoid certain combinations of consonants? If so, could this be taken to be a property of Proto-Atlantic which is preserved in the daughter languages? In this case, the shunned combinations could provide a solid argument in favor of the existence of the Atlantic group itself, for which no shared linguistic trait has been offered as evidence of the alleged genetic sub-branch of Niger-Congo. In addition, if valid, consonant distribution patterns of this sort might furnish very interesting pathways for the reconstruction and comparision of Proto-Atlantic with other subgroups of the Niger-Congo macro-family. In this case the appropriate research strategy would be to establish the precise consonant combinations favored or disfavored in languages from each of the other sub-groups. To do so, one would have to transform any intuitive impression into numeric values.[3]

Let us return to the "simplified" roots, where each consonant is represented by the symbol of its class. For each biconsonantal root, there is an initial consonant (C_1) and a non-initial consonant (C_2). For longer roots, let's say of the form CVCVC, the medial consonant is C_2 with respect to the initial consonant, but it is C_1 with respect to the final consonant. In our statistical counts, the lexicon is thus broken up into C_1-V-C_2 sequences, where the medial consonant of a triconsonantal root is calculated both with respect to the consonant which precedes and the consonant which follows it. For sequences including consonant clusters, as for example CVCCVC, the cutting was CVC-CVC; when two adjacent consonants are of the same class, as for example CVPPVC, then it is equivalent to the CVPVC case, the P being final for the first sequence and initial for the second one.

As an illustration, consider the consonant system of Balanta, displayed by place and manner of articulation in Table 2.

f	t, th	c, s	k, h
b	d	j	g
gb			
w	l, r	y	
m	n	ɲ	ŋ
mf	nt, nth	ns	
mb	nd	nj	ŋg
ŋgb			

Table 2. The Balanta Consonant System

As seen, there are eight labial consonants /f, b, gb, w, m, mf, mb, ŋgb/, nine dental consonants /t, th, d, l, r, n, nt, nth, nd/, seven palatal consonants /c, s, j, y, ɲ, ns, nj/ and five velar consonants /k, h, g, ŋ, ŋg/, which we can symbolize by P, T, C, and K, respectively.[4] Table 3 presents the measured frequencies of each place of articulation

[3] For an application of statistical methods to comparative studies, see Pozdniakov (1991).

[4] /th/ and /nth/ are interdentals. /s/ is treated as palatal, as it frequently occupies the " [ʃ] slot", and labiovelars are treated as labials. These choices have been suggested by our knowledge of the Atlantic languages, where for example /s/ is associated with /c/ in all the languages showing consonant mutation, as Fula, Sereer, the Tenda cluster, Wolof, etc. We

from a Balanta lexicon of 766 entries containing 904 C_1VC_2 sequences[5] (Ndiaye-Corréard 1970):

	P	T	C	K
C_1	26.8	31.7	23.7	17.8
C_2	20.5	48.7	14.8	16.0

Table 3. Balanta: Observed frequencies (O), in %

The next step in the procedure is to calculate the theoretical frequencies of the different combinations. The theoretical frequency of any of the 16 C_1-V-C_2 combinations is obtained by multiplying the absolute frequency of the C_1 consonant by the absolute frequency of the C_2 consonant. The theoretical frequency thereby obtained assumes an absence of correlation between the quality of the C_1 and that of the C_2. Table 4 furnishes the theoretical frequencies of the 16 combinations for Balanta:

		C_2			
		P	T	C	K
C_1	P	5.5	13.1	4.0	4.3
	T	6.5	15.4	4.7	5.1
	C	4.9	11.5	3.5	3.8
	K	3.6	8.7	2.6	2.8

Table 4. Balanta: Expected frequencies (E), in %

To illustrate, consider the example of K-P, that is, cases where any velar is followed by any labial consonant. The frequency of C_1 K is 17.8 %, and the frequency of C_2 P is 20.5%. The frequency of the combination K-P is therefore theoretically 17.8% x 20.5%, or 3.6%. Based on the 904 sequences examined, this means that one should find 904 x 3.4%, or 33 sequences of the form KVP. Among the 766 entries in the Balanta lexicon, one does in fact find 34 sequences of this form. We can therefore consider this distribution as "normal" (i.e. conforming to the anticipated theoretical frequency).

On the other hand, the combination K-K, for which we should find 17.8% x 16.0% x 904 = 26 sequences, attests only 9 such forms. For each language, one thus compares the theoretical or expected (E) frequency with the actual or observed (O) frequency of each combination. If a correlation exists between the qualities of C_1 and C_2, this should be manifested by a significant discrepancy between the E and O values. Thus, taking the example of the actual K-K sequences in Balanta, the discrepancy is -65% with respect to the theoretical frequency. Table 5 presents the E/O discrepancies for C_1-C_2 in Balanta.

		C_2			
		P	T	C	K
C_1	P	–57,6	+30,7	+19,9	–38,2
	T	+22,6	–22,0	+1,1	+36,9
	C	+32,4	–20,3	–24,3	+42,8
	K	+3,2	+20,0	+0,6	–65,1

Table 5. Balanta: 100(O–E)/E*

chose not to change these while computing data from other languages. Other possibilities would have been to treat /s/ as a dental, or labiovelars as velars, or both. This would have slightly changed the figures, but not the tendencies. Moreover, as will be shown soon, dentals share some statistical features with palatals while labials share some features with velars.

[5] The number of sequences used for calculating the tables is indicated by the mention *n = xx* in all the tables from now.

By convention and for purposes of readability, we adopt the following procedure in presenting our results:

(i) A discrepancy whose absolute value is less than 15% is considered to be non-significant and is not noted.

(ii) A discrepancy whose absolute value is between 15% and 30% is noted by a + or - sign.

(iii) A discrepancy whose absolute value is greater than 30% is noted by a double + + or – – sign.[6]

The actual values, i.e. the observed number of each combination, are given in Table 32 as an appendix at the end of the paper.

Following these conventions the percentages in Table 5 produce the values in Table 6.

		C_2			
		P	T	C	K
C_1	P	– –	+ +	+	– –
	T	+	–		+ +
	C	+ +	–	–	+ +
	K		+		– –

Table 6. Results of Table 5 in terms of +/– categories

The presentation in Table 6 has the advantage of nicely exposing the positive and negative tendencies. Thus, as indicated by the minuses along the descending diagonal, Balanta systematically underrepresents combinations of consonants at the same place of articulation. Not only is it the case that velars rarely combine (as seen in the discussion of K-K above), but the same is true of labials, palatals and to a somewhat lesser extent dentals. The systematic nature of the distribution observed in Table 6 (and elsewhere to follow) is a good indicator of the validity of the general approach and of the specific method. Such results in fact encourage us to seek other distributional regularities.

Before moving on to consider other languages, we should take note of two other observations that can be made on the basis of Table 6. First, the presence of negative discrepancies must be compensated by positive discrepancies. By definition, the total sum of the discrepancies with respect to the norm must be zero. The negative discrepancies are almost exclusively due to the principle of SPA: Balanta disfavors sequences where C_1 and C_2 are made at the same place of articulation. On the other hand, the positive discrepancies do not seem to be principled. We thus do not see the relationship between the fact that C_1 palatals preferentially combine with C_2 labials and velars and the fact that C_1 labials more frequently combine with C_2 palatals. In other words, while the distribution of the minuses is (relatively) regular and systematic, the distribution of the plusses is not.

A second observation that can be made from Table 6 is that from the statistically point of view, the four consonant classes P, T, C et K can be grouped into two "superclasses": P and K vs. T and C. Within roots, not only do palatal consonants show a statistical tendency to not combine with another palatal, but also not with a dental. Similarly, labial consonants tend not to combine with other labials, but also not with velars. We will refer to the two superclasses as "peripheral" (P, K) and "medial" (T, C). The peripheral/medial opposition corresponds exactly to the grave/diffuse distinction of Jakobson *et al.* (1952) and inversely to the coronal/non-coronal opposition of generative phonology (Clements & Hume 1995).

[6] For ease of readability we do not use the χ^2 test. As another advantage, the method used here preserves the direction of deviation with respect to the norm, which χ^2 does not.

The two superclasses behave as basic classes in that consonants from within each set rarely combine with each other. However, of equal significance is the fact that the lack of combinations within a superclass corresponds to an excess of combinations between the superclasses, as summarized in Table 7.

	Peripheral	Medial
Peripheral	–	+
Medial	+	–

Table 7. Superclasses

Because of these new groupings, we propose to modify the order of presentation within the tables. Instead of the articulatory order PTCK, we shall henceforth adopt the order PKTC, as shown for Balanta in Table 8.

		C_2			
		P	K	T	C
C_1	P	– –	– –	+ +	+
	K		– –	+	
	T	+	+ +	–	
	C	+ +	+ +	–	–

Table 8: Balanta: PKTC order

As seen by the bold borders, we can now distinguish four quadrants in these tables: The upper left and lower right quadrants enclose combinations where C_1 and C_2 belong to the same superclass. As seen, all of the minus signs fall within these quadrants. The lower left and upper right quadrants indicate combinations where C_1 and C_2 belong to different super-classes. All of the plus signs fall within these quadrants. Finally, the C_1 and C_2 which combine in the cells along the descending diagonal belong to the same class (place of articulation). Here and in the subsequent tables, these cells are shaded.

The Balanta facts have served as an illustration in presenting the methodology and an initial set of results for comparison. In the following sections we shall see that SPA by class and superclass is widespread in Africa and beyond.

2. SPA in Atlantic

As mentioned, our initial interest was both diachronic and Atlantic-specific: In all of the Atlantic languages we have examined, one finds the same distributional tendencies concerning the classes P, T, K, C, as well as the peripheral and medial superclasses. We present the results from 12 Atlantic languages in Table 9, where the cells along the descending diagonal representing same place combinations are shaded[7]:

Fula (Labatut 1994; n = 672)

	P	K	T	C
P	– –	– –	+	
K	– –	– –		+ +
T	+ +	+ +	–	–
C	+	+ +		– –

Palor (d'Alton 1987; n = 2,116)

	P	K	T	C
P	– –	–	+	+
K	–	– –	+ +	
T	+ +	+ +	– –	
C		+ +	–	–

[7] From this point on, the number of CVC sequences computed for each language is given above each table as *n = xx*.

Wolof (Fal et al 1990; n = 8,456)

	P	K	T	C
P	– –	–	+	+
K		– –	+	+
T	+	+ +	–	
C	+	+		–

Nyun-Buy (Lespinay 1991; n = 4,428)

	P	K	T	C
P	– –			+ +
K	–	– –	+ +	+
T	+	+ +	–	– –
C	+ +		–	

Jaad (Ducos 1971; n = 1,200)

	P	K	T	C
P	– –	– –		+ +
K		– –		+
T	+ +	+ +	–	– –
C	+ +	+		– –

Balanta (Ndiaye-Corréard 1970; n = 904)

	P	K	T	C
P	– –	– –	+ +	+
K		– –	+	
T	+	+ +	–	
C	+ +	+ +	–	–

Joola Kwaatay (Payne 1992; n = 2,183)

	P	K	T	C
P	– –	–	+	+
K	–	– –	+	
T	+ +	+	–	
C	+	+	–	

Manjaku (Buis 1990; n = 3,145)

	P	K	T	C
P	– –	– –	+	+
K	–	–		+
T	+ +	+	–	
C		+		–

Bijogo (Segerer 2002; n = 1,499)

	P	K	T	C
P	– –	–	+	+ +
K		– –	+ +	
T	+ +	+ +	–	–
C	+ +			– –

Sua (Segerer 1998; n = 495)

	P	K	T	C
P	–	– –		+ +
K	–	–	+	n.s.[8]
T	+	+		–
C		+		

Bullom (Nyländer 1814; n = 827)

	P	K	T	C
P		–		+ +
K		– –	+ +	*n.s.*
T	+	+ +	–	– –
C				

Kisi (Childs 2000; n = 2,981)

	P	K	T	C
P	–	–	+	+
K	–	–	+ +	
T	+ +	+	–	– –
C		+	–	

Table 9. CVC combinations in 12 Atlantic languages

The shaded cells along the descending diagonal are almost all characterized by one or two minus signs. This corresponds to the tendency for roots to avoid C_1-C_2 sequences made at the same (or similar) place of articulation. The tables show that this tendency also affects consonants from the same superclass: Peripheral consonants tend not to combine, as do medial consonants tend to avoid one another. The P/K and T/C groupings, which one might have considered specific to our discussion of Balanta in §2, are relevant in all of the languages examined, without exception.

Let us separately examine the two superclass diagonals in the above tables, each one consisting of two quadrants. One, the "grey super-diagonal", represents the combination of consonants of the same superclass (including the shaded descending diagonal). The other, the "white super-diagonal," represents the combination of consonants of different superclasses. Given the 12 languages examined, for each of the combinations peripheral-peripheral, peripheral-medial, medial-peripheral, and medial-medial, there are 12 x 4 = 48 cells to fill (since each combination of two superclasses corresponds to four combinations of place). The result is striking: As seen in Table 10, all of the minus signs are concentrated in the grey super-diagonal:

[8] The abbreviation *n.s.* (non-significant) indicates the value of the norm (E) is too low to have a statistical value. If the expected amount of a given combination is 2 and we find three examples of this combination, it will make a 50% positive deviation. We consider that in such cases the deviation is not relevant, because the influence of chance is too big . We arbitrarily set the minimal value for the norm at 10.

	Peripheral	Medial
Peripheral	41 / 48	0 / 48
Medial	0 / 48	30 / 48

Table 10. Number of minusses in each quadrant

In other words, the number of combinations of consonants from the same superclass is always less than the norm, and the number of combinations of consonants from different superclasses is never less than the norm.

This very marked tendency found in all branches of the Atlantic group motivated us to postulate a similar distribution in Proto-Atlantic. As Atlanticists, the possibility of such a reconstruction was good news to us for two reasons: First, it provided us an element of proof in establishing the reality of the Atlantic group. As already mentioned, no one up to this point had cited a single linguistic trait found only in the Atlantic group. Second, the reconstruction of SPA at the Proto-Atlantic level seemed to open new perspectives in seeking regular correspondences with languages within other sub-branches of Niger-Congo: If SPA were the result of a Proto-Atlantic innovation involving place dissimilation, then it might be that Atlantic labials in a certain context correspond to Niger-Congo dentals, or that Atlantic velars correspond to Niger-Congo palatals. As seen in the following sections, what has instead turned out to be "bad" news for Proto-Atlantic has wider consequences for the study of language in general.

3. SPA in Niger-Congo

If it is reasonable to postulate that Proto-Atlantic innovated SPA, it should be the case that, statistically, other Niger-Congo sub-groups do not exhibit the same systematic restrictions in their consonant distributions. In other words, if SPA is really an Atlantic innovation, it should be absent in other Niger-Congo subgroups. This was our belief, but we were wrong. In this section we examine several other sub-branches of the Niger-Congo phyllum for which we have reconstructions.

Table 11 presents the results obtained by mapping out Moñino's (1995) lexical reconstructions of Proto-Gbaya.[9]

		C_2			
		P	K	T	C
C_1	P	– –	– –	+	+ +
	K		– –		+ +
	T	+ +	+ +	–	– –
	C		+ +		n.s.

Table 11: Proto-Gbaya (Moñino 1995; n = 761)

As seen, the situation is similar to what we saw in Atlantic: All of the same-place combinations are avoided, and the remaining minus signs concern combinations within the same superclass. Similar results are found in Proto-Ijo:[10]

[9] The Gbaya languages which are spoken in Central Africa (Cameroun, Central African Republic) constitute a branch of Niger-Congo, perhaps at the same level as Adamawa or Gur.

[10] Proto-Ijo (Nigeria), which has been reconstructed by Williamson (2004), constitutes one of the highest branches of the Niger-Congo phylum.

		C_2			
		P	K	T	C
C_1	P	– –		+	
	K	–	– –	+	n.s.
	T	+ +	+ +	– –	n.s.
	C	+	n.s.	–	n.s.

Table 12: Proto-Ijo (Williamson 2004; n = 509)

Again, combinations of homorganic consonants are avoided. With distant Proto-Gbaya and Proto-Ijo joining Proto-Atlantic, the alleged "Atlantic dissimilation" is obviously not an isolated fact. The probability of systematic SPA occurring by inheritance in three Niger-Congo branches is extremely low. At this point we are conditioned to expecting the same grey diagonal minuses in other branches. Table 13 shows that Proto-Mande, another early off-shoot of Niger-Congo, does not disappoint:[11]

		C_2			
		P	K	T	C
C_1	P	– –	–		+ +
	K	n.s.	– –	+	n.s.
	T	n.s.	+ +		n.s.
	C	n.s.	+ +	–	n.s.

Table 13: Proto-Mande (Vydrine 2004; n = 511)

As in previously examined cases, all of the minuses are concentrated within the grey super-diagonal, and the only double minuses are in the grey cells within the narrow diagonal. Thus, as far back as Proto-Mande we see a clear avoidance of consonant combinations at the same place of articulation and, to a lesser extent, consonants belonging to the same superclass. In addition, one observes that within the shaded cells along the narrow diagonal, SPA is stronger among peripheral vs. medial consonants. In fact, this observation is valid for all of the languages described up to now.

The only languages which show a certain deviation from the reported regularities are the Bantu languages, whose combinatorial properties are indicated in Table 14.

		C_2			
		P	K	T	C
C_1	P	–		+	
	K		– –	+	
	T	+	+		– –
	C	+	+	– –	+ +

Table 14: Proto-Bantu (Bantu Lexical Reconstructions 1998; n = 12,426)

Here for the first time, a grey cell in the table has a '+ +'. It appears that Proto-Bantu had access to a disporportionate number of palatal consonant sequences. Since this has to do with medial consonants, another tendency evoked above remains valid: Peripheral consonants avoid each other more. An examination of individual Bantu languages from different zones (Guthrie 1967-71) shows that the Proto-Bantu situation is well-reflected in present-day daughter languages:

[11] The calculations in Table 13 are based on the Proto-Mande reconstructions presented by Valentin Vydrine at the Workshop on Proto-Niger-Congo, Paris 2004.

Swahili (zone G; n = 1,481)

	P	K	T	C
P	– –	+		
K		– –	+ +	+
T	+		–	–
C	+		– –	

Mpongwe (zone B; n = 3,506)

	P	K	T	C
P			+	+
K		–	+	
T	+			– –
C		+	–	+ +

Bemba (zone M; n = 10,653)

	P	K	T	C
P	–		+	
K		– –	+	+
T	+	+		–
C		+	–	+ +

Kiga-Nkore (zone J; n = 17,944)

	P	K	T	C
P	–			
K		–	+	–
T		+		–
C	+		–	+

Table 15. Four individual Bantu languages of four different zones

One does not have to be a specialist in Bantu historical linguistics to assume that the unusual statistical distribution of palatal consonants in the reconstructed roots as well as in present-day languages reflects a Proto-Bantu innovation with respect to Proto-Niger-Congo. In fact, the Bantu languages are the only ones which show any tendency for C_1 and C_2 consonants to agree in place of articulation—and, except for Swahili P-K, only among palatals.

4. SPA as an African areal feature?

The Bantu deviation just discussed should not hide the fact that SPA represents a formal characteristic of the entire Niger-Congo family. The question addressed in this section is whether SPA is specific to Niger-Congo or whether it is an African areal feature. If the former, then languages from other African phyla should have behaviors which are significantly different from those just seen in Niger-Congo. We begin with Sara-Kaba-Na, a Nilo-Saharan language of the Sara-Bongo-Bagirmi branch:

		C_2			
		P	K	T	C
C_1	P	– –	– –	+ +	
	K		– –		+
	T	+	+ +	–	
	C		+ +		– –

Table 16. Sara-Kaba-Na (Danay K., M. Makode et al. 1986; n = 3,300)

As seen, the distributions are absolutely comparable to what we have thus far observed in Niger-Congo: The grey diagonal is entirely filled with minuses, and we find no minus outside the grey super-diagonal. If these distributions could be shown to be a valid genetic marker, they would have something to offer to those who favor a union of the Niger-Congo and Nilo-Saharan families into a even larger macro-family (but cf. §5).

Let us therefore take a language of Africa which we know not to be involved in this hypothesis: Based on its unique genetic source and relative isolation from the continent, Malagasy, an Austronesian language, would not be expected to share linguistic properties with Niger-Congo or Nilo-Saharan languages. One nevertheless clearly sees in Table 17 the same discrepancies with respect to the norm to which we have become accustomed:

		C_2			
		P	K	T	C
C_1	P	– –		+	+
	K		– –		+
	T	+ +		–	–
	C		+ +	–	– –

Table 17. Malagasy, Sakalava (Lacroix, 2001; n = 1,944)

What better example could we find of a typical Niger-Congo, even Atlantic distribution? All of the minuses are in the grey super-diagonal, and all of the plusses are in the white super-diagonal. At this point of the investigation, we begin to have reasons that we are dealing with a more general phenomenon which goes beyond the boundaries of genetic divisions. Could SPA be an African areal feature?

The examination of African languages would not be complete without representation from the Afro-Asiatic phylum, here represented by the Chadic subgroup.[12] In this connections we have only tested two lexicons: the tentative Proto-Chadic reconstructions of Jungraithmayr & Ibrizsimow (1994) and the lexicon of the Ader dialect of Hausa dialect (Caron 1991). The results are presented in Table 18.

Proto-Chadic (n = 1,306)

	P	K	T	C
P	– –		+	+ +
K		– –	+	+
T	+ +	+ +	– –	
C	+ +	+	–	– –

Hausa (n = 3,880)

	P	K	T	C
P	– –			+ +
K	–	–	+	
T	+	+		–
C	+ +		–	

Table 18. Chadic and Hausa (Ader dialect)

In the Proto-Chadic corpus, the shaded diagonal is even more marked than in Atlantic or other Niger-Congo languages. Recall that the '– –' sign indicates that the number of combinations is at least 30% below the norm calculated according to the percentages observed independently in each position. Here the situation is uniform for all four combinations of homorganic consonants. In Hausa the tendency is a little less strong, but it is not contradictory, since all the minuses remain in the grey super-diagonal. As in Niger-Congo, the tendency is greater for peripheral than for medial consonants.

One might perhaps be less surprised that SPA is in full force within Chadic than in the other language families we have examined. Chadic is a branch of the Afro-Asiatic macro-family to which Arabic, Hebrew and the Semitic subgroup also belong. As stated in our abstract and in §0, the avoidance of combinations of similar consonants has been noted in these languages for some time, and with Chadic we can extrapolate perhaps to the level of Afro-Asiatic itself.

Since Semitic takes us marginally outside of Africa, the possibility that SPA is an African areal feature is somewhat weakened. The final blow comes in the next section, where we demonstrate that SPA is a linguistic universal.

5. SPA as a linguistic universal

Before presenting evidence for SPA from outside the African continent, where we have less expertise, we wish to comment again on the steps that have been involved in conducting this study. We reiterate that we first discovered the SPA phenomenon in the Atlantic languages. As specialists of these languages we have developed tools for treating

[12] Unfortunately we have not been able to study any language from Khoisan, the fourth African macro-phylum.

important corpora without making too many methodological errors. However, we were surprised to see the SPA phenomenon so distinctly manifested. It is clear that we do not have the same expertise to treat the data from other families. Despite this, and despite any possible biases, we have noted the same tendency working on lexicons which have not been subject to the kind of phonological and morphological analyses that should logically precede this kind of calculation.

To illustrate this, we conducted an experiment based on Fula. The whole corpus has 1153 items (1651 CVC sequences). However, out of these 1153 words, there are only 643 primary lexical stems (672 CVC sequences), the others being derived. So we calculated the tables for both corpora:

"clean" corpus, n = 672

	P	K	T	C
P	– –	– –	+	
K	– –	– –		+ +
T	+ +	+ +	–	–
C	+	+ +		– –

"raw" corpus, n = 1,651

	P	K	T	C
P	–	– –		+
K	– –	–		+ +
T				–
C	+ +	+ +		– –

Table 19. Fula: difference between "clean" and "raw" corpora

We can see that the general tendency is the same in both cases, even if the details are different: All the minusses are in the descending diagonal and all the plusses are in the ascending one. In addition, we can see that SPA is in every case more important for peripheral consonants than for medial ones. This experiment shows that even with no expertise on a given language, we are allowed to make statistical measures on this language.

In each of the languages examined up to now, the same tendencies have been at work. The African languages presented in the preceding sections belong to four different families, but are still tied by geography. The question that now arises is: What happens outside of Africa? It seems that even in Indo-European statistical skewings in the combination of C_1-C_2 consonant sequences have not been very appreciated. Table 20 presents our findings for Proto-Indo-European (PIE), based on the Starling Data Base of Starostin (1998-2005).

		C_2			
		P	K	T	C
C_1	P	– –		+	
	K		– –	+	
	T	+ +	+ +	– –	
	C	+	+		–

Table 20. Proto-Indo-European (n = 3,085)

Indo-European is without a doubt the most studied language family, and that for more than two centuries. As seen in Table 20, the same SPA tendency is observed in PIE reconstructions. We note again that combinations of consonants at the same place of articulation exist in most, if not all languages, and that they are not necessarily rare. Some examples from Indo-European taken from Starostin (1998-2005) include:

(i) labial–labial: **pib* 'to drink' ; **paw* 'few, small' ; **bhebhr-u-* 'bear'
(ii) dental–dental: **tal-/-e-* 'earth, ground' ; **del* 'long' ; **nan-/*nen-* 'mother, nurse'
(iii) palatal–palatal: **yes* 'to boil'
(iv) velar-velar: **(s)kek-* 'hair, beard' ; **koks-* 'armpit'

It is likewise quite easy to find examples of consonant sequences at the same place of articulation in African languages in general and in Atlantic in particular. Often these are

words of great frequency of usage, e.g. Wolof *bopp* ‘head’, *sàcc* ‘steal’. From such examples one might easily conclude that there are no constraints on transvocalic C_1-C_2 sequences. As this study has documented, this would be an error. As we have shown, *statistically*, these combinations are relatively rare.

Although the notion of a Proto-Nostratic existing at a considerably greater time depth than Proto-Indo-European is quite controversial, Table 21 shows that SPA is observed in the reconstructions proposed in the etymological dictionary of Illych-Svytych (1971-1984):

		C_2			
		P	K	T	C
C_1	P	– –	–	+ +	–
	K		– –		+ +
	T	n.s.	+ +	– –	– –
	C	n.s.	n.s.		n.s.

Table 21. Proto-Nostratic (n = 318)

Since Altaic is one of the branches of Nostratic, it is not surprising to find SPA effects in the languages of that family. Table 22 presents the facts of “Classical” Mongolian:

		C_2			
		P	K	T	C
C_1	P	– –	–	+	+ +
	K		– –	+ +	
	T		+ +	– –	
	C	+			–

Table 22. Classical Mongolian (n = 66,407[13])

In this table the tendencies are more marked than ever: Not only do the shaded cells contain 7 minuses out of 8 possible, but also the cells of the inverse diagonal contain 7 out of 8 possible plusses. (Only one minus and one plus occur outside these “narrow” diagonals, but both are found within expected quadrants.) The distribution of these plusses strikingly suggest that the heterorganic combinations P-C, C-P, T-K, and K-T are strongly favored in Classical Mongolian. We have already remarked that the two diagonals do not have the same status: While there is a tendency to find the most minuses along the descending (shaded) diagonal in all of our tables, there does not appear to be a correspondency tendency for the greatest number of plusses to congregate along the inverse diagonal. Rather, these plusses appear randomly distributed within the cells of the lower left and upper right quadrants.

While the other languages examined are not systematic in their preferences for certain heterorganic sequences, the question naturally arises as to whether the distribution of the plusses in Classical Mongolian is in fact principled. We summarize the relevant facts as follows, where “>>” means ‘is preferred over’:

(i) P-C >> P-T
(ii) T-K >> T-P
(iii) C-P >> C-K
(iv) K-T >> K-C

As seen, combinations of labials and palatals (P, C) are preferred over combinations of labials and dentals (P, T), and combinations of dentals and velars (T, K) are preferred over

[13] The Mongolian data were extracted from an on-line dictionary of over 25,000 entries, available at the following address: http://membres.lycos.fr/brunogml/sub/corps.htm.

combinations of dentals and labials (T, P). Up until now we only recognized peripheral (P, K) and medial (T, C) superclasses, which group together the most similar places of articulation. The commonality of the two places of articulation in each superclass can be defined either in terms of their shared acoustic properties (grave vs. acute) or their shared articulatory (non-)involvement of the front of the tongue (coronal vs. non-coronal). Mongolian now suggests that anterior consonants (P, T) and posterior consonants (C, K) also share a property, which corresponds roughly to [±high] (raising of the body of tongue) in the Chomsky & Halle (1968) distinctive feature framework. In this framework the four places of articulation would have the feature values in Table 23.

	P	T	C	K
coronal	–	+	+	–
high	–	–	+	+

Table 23. Shared features among P, T, C, K

Approached in these terms, we see that two groupings do not share either feature: P, C and T, K. It could therefore be that Classical Mongolian has the flip-side of SPA, namely the favoring of the most dissimilar consonant sequences. We note that Classical Mongolian is the only language in our study which has front-back vowel harmony, which may turn out to be a relevant factor, hence worthy of further study.[14]

To summarize thus far, we have seen that SPA effects are widespread in the world's languages. Even if we have not tested all of the languages or language families of the world, there is reason to believe that we are dealing with a universal phenomenon. What would it take to be even more convincing, specifically to rule out any possibility of an Afro-Eurasian genetic or contact phenomenon? A genetically isolated language? A recently formed language, e.g. a pidgin? A language belonging to more exotic language families (Australian, American Indian)? Table 24 presents an example of each of these:[15]

Basque, Euskara (n = 3,140)

	P	K	T	C
P	– –			
K	– –	– –	+	
T	+	+ +	–	+
C	+ +			– –

Pidgin English, Port-Moresby (n = 2,215)

	P	K	T	C
P	– –		+	
K		– –		
T	+ +	+	–	
C		+		

Quechua (n = 5,254)

	P	K	T	C
P	– –	–	+ +	
K		– –		
T	+	+ +		–
C	+	+ +	–	

Kamilaroi, Australia (n = 980)

	P	K	T	C
P	– –	+ +		
K	+ +	– –		
T	+			+
C		–	+	– –

Table 24. Other languages

In this arbitrary sample of four languages we note no contradiction with the tendencies previously seen. Of the 16 shaded cells in Table 24, 12 contain one or two minuses, and none contains any plusses. In contrast, the tendency is less clear concerning the "wide" diagonal, i.e. combinations within the same superclass. This is particularly striking in the case of Kamilaroi, where P-K, K-P, T-C and C-T are overrepresented. As seen, the minusses found along the grey diagonal are unexpectedly compensated by plusses in the white cells of the upper left and lower-right quadrants. This case is unique among our

[14] Many of the African languages cited have either ATR or height harmony which may be expected to interact less with SPA than front-back vowel harmony. More languages having the latter, as well as rounding harmony, need to be investigated (e.g. Turkish).

[15] All the data for theses four languages were found on the internet. The links are given at the end of the *References* section.

sample. Perhaps what we can say is that even if every language shows the effects of SPA, each language preserves its own originality. Out of the 31 tables presented above, no two are exactly identical. Evaluating the significance of the individual differences is of course a task reserved for specialists of each language and language family.

6. Discussion

The preceding sections have clearly established that SPA is a likely universal property of human language. We are aware that other studies have been concerned with the tendency of like features or segments to resist repetition or be kept at a distance from one another. Within non-linear phonology, the Obligatory Contour Principle (OCP) is often cited as a universal tendency:

(1) Adjacent identical elements are prohibited (McCarthy 1986:208).

Various authors have invoked the OCP or related principles under different names to account for a variety of phenomena that minimize the same or similar elements (consonants, vowels, tones, whole syllables, etc.). The following quote from Tang (2000:34) succinctly references much of this work:

> "In the literature, the principle in (1) has been called the *obligatory contour principle* (OCP, Leben 1973; Goldsmith 1976; McCarthy 1986), the *repeated morph constraint* (Menn and MacWhinney 1984), ANTIHOMOPHONY (Golston (1995), *REPEAT (Yip i.p.), and IDAVOID (Brentari 1998). The effects of this principle not only can be observed in autosegmental phonology and feature geometry (Leben 1973; Goldsmith 1976; McCarthy 1986; Myers 1987; Yip 1988b; Pierrehumbert 1993, among many others) but also can be found in morphology (Stemberger 1981; Menn and MacWhinney 1984; Mohanan 1994; Golston 1995; Yip 1995, 1998; Brentari 1998, among many others)."

Within phonology, SPA effects of the type described in this paper have long been observed within Semitic languages and continue to be the subject of study, especially as concerns Arabic (Frisch, Pierrehumbert & Broe 2004) and Hebrew (Berent & Shimron (2003). While the SPA effects have been presented as static distributional tendencies, the diachronic process of consonant dissimilation has received considerable attention for over a century, mostly notably in Grammont (1895). The question we would like to raise in this section is whether the SPA effects we have reported in tabular form are one of the manifestations of the OCP. As attractive as this may seem, our approach and findings differ from some of the above work in at least three ways:

(i) In spite of a few exceptions (Yip 1995, Frisch *et al.* 1996, MacEachern 1999, Berent & Shimron 2003, Coetzee & Pater 2006) the universal OCP mostly concerns *identical* elements. In our case we have dealt not only with restrictions on combinations of identical, but also similar elements: We have been concerned with restricted sequences of consonants made in a same or similar place of articulation. We have also shown that the restrictions hold not only of exact homorganic consonants, but also of consonants that belong to the same "superclass". In this connection, we have seen the SPA at work within both the medial and peripheral superclasses, although the tendency has a greater effect among peripheral consonants.

(ii) The results obtained for Semitic exploit the fact that in these languages consonantal roots have a concrete reality which can be observed in their templatic morphology. Their isolability makes the calculations relatively easy. Our measurements have involved languages where the notion of 'consonantal root' is rarely justified. We have even

evaluated corpora with no preliminary knowledge of the morphological structures of the languages in question. Despite this, the SPA effects were still evident.

(iii) Whereas recent discussions of SPA, especially those based on Semitic, seek above all to discover the nature of this phenomenon in synchrony,[16] for us the phenomenon has great diachronic consequences for the comparative method. If consonant combinations have to satisfy a principle of equilibrium, the phonetic changes that affect consonants will presumably be in part conditioned by SPA. Thus, alongside the two major sources of language change, regular sound change and analogical change, a third factor must be at play which we can call 'dissimilation consonantique'. Within this category we can classify the examples given by Greenberg (1968:107-108) such as Latin *arbor* > Spanish *arbol* or Latin *anima* > **anma* > Spanish *alma*. Here we have neither a regular sound change (*r > l or *n > l) nor analogical change, but rather "sporadic" dissimilations: of two *r*'s in the first case, of two nasals in the second. Greenberg formulates this tendency only for sonorants (liquids and nasals) and s(h)ibilants, thus for specific manners of articulation. We have tried to show the importance of similarity avoidance for place of articulation.

Since SPA is general in language, we should expect to find processes that affect C_1VC_2 sequences where C_1 and C_2 belong to the same class or superclass with respect to place. Logical possibilities include one consonant dissimilation from the other in place, or dropping out under identity. Another possibility is that lexical items that repeat the same place of articulation may be disfavored and drop out—or may not have been formed in the first place. On the other side of the equation, when we find a language which violates SPA in an unexpected way, we might conclude that the language in question has undergone a specific diachronic change to produce the unusual situation. For example, in Bantu we observed an excess of C-C (palatal-palatal) combinations. Since Bantu diverges significantly from the norm in this respect, it behooves the Bantuist scholar to seek a diachronic explanation. From a wider comparative point of view, this unique distribution leads one to hypothesize that the homorganic C-C sequences of Proto-Bantu must correspond to other consonant combinations in other groups of Niger-Congo.

The Bantu example might give the impression that only such an "anomaly" can be exploited for comparative purposes. However, although all of the languages presented here show the same general SPA tendency, they all differ in their statistical details, which may therefore furnish precious indices for comparison.

As we indicated in §5, there may also be an important interest in closely studying the possible relationship between SPA and the inverse tendency of vowel harmony. Vowel harmony is of course a quite different phenomenon operating synchronically and productively, contrary to consonantal incompatibility. As we hypothesized in the case of Classical Mongolian (Table 22), the tendency for consonant place to be dissimilar may be more accentuated in languages with front-back vowel harmony. If the distribution of plusses in the inverse diagonal of Table 22 is not fortuitous or isolated, this could mean that there is a tendency for an equilibrium to be established between the two sub-systems (vowels and consonants) within the phonological structure of the word.

There is one limitation on SPA that we have not yet addressed. Contrasting with the very clear tendency to avoid successive consonants of similar place, the Atlantic languages shows an inverse tendency which at first appears contradictory: In Atlantic, as in other languages of the world, *identical* consonants combine easily, especially in certain lexical subclasses, e.g. ideophones, intensifier adverbs, and other iconic words. One of the

16 Cf. Frisch *et al.*, op. cit., especially §4.1 "The Psychological Reality of OCP-Place" (p. 210).

possible sources of combinations of identical consonants is reduplication, which is often associated with the expression of intensity. However, the statistical results presented here in support of SPA are robust *despite* these well-known cases of identical consonant combination. It is not impossible that these two opposing tendencies are interrelated.

To test the effect of identical C_1-C_2 consonants on our results, we did a more fine statistical count on Wolof which distinguished between sequences of identical vs. non-identical homorganic consonants. This further study yielded the following results:

(i) If one takes into account the relative frequencies of the different consonants in the dictionary, there is no general tendency to combine identical consonants, with the exception of two special cases: (a) Combinations of identical nasal or prenasalized consonants, especially *mb-mb, nd-nd, m-m,* and *n-n,* are considerably more frequent than expected; (b) The sequences *f-f* and *c-c* are particularly frequent.

(ii) Besides the combinations whose frequency exceeds or corresponds to the norm, two combinations of identical consonants have a frequency below the norm: *r-r* and *t-t.*

(iii) An analysis of lists of words which present sequences of identical consonants shows that these words are not generally formed by lexical reduplication. They are, however, often formed by a sort of "grammatical" reduplication. In Wolof, for each noun class there is a series of determiners of the form *CooCV* where C is the consonant of the noun class and V is a "deictic" vowel (*i* for 'near', *a* for 'far', *u* for 'unmarked'). Other forms exist which are accompanied by a particle with an emphatic value (*i* or *le*), which increases the number of words containing a sequence of two identical consonants. For example, for the "M class", we find:

muus mi	'this cat'
muus moomu	'this cat (in question)'
muus moomule	'idem (emphatic)'
muus mooma	'that cat (of which you had spoken and which is not present)'
muus moomale	'idem (emphatic)'
muus moomee	'idem (emphatic)' (< **mooma* + *i*)
muus moomii	'this cat here (of which you had spoken, emphatic)'

Besides the above, the consonant *n* is repeated in the lexical base meaning 'other', e.g. *m-eneen* in the M class. The presence of these multiple series increases the frequency of combinations of identical consonants.

In the above examples, the repetition of identical consonants is associated with the morphology of the language. They are thus exempt from any possible phonetic or phonological motivation for SPA. If we remove all of the words that have a sequence of identical consonants from the wordlists, the tendency for successive consonants to be of different place is of course enhanced. This is seen in Table 25 which compares the percentages of each combination with vs. without identical C_1=C_2 sequences:

with	P	K	T	C
P	-54	-17	+23	+17
K	-9	-69	+27	+15
T	+27	+31	-22	-4
C	+26	+30	-14	-26

without	P	K	T	C
P	-74	-15	+28	+23
K	-6	-83	+30	+20
T	+35	+37	-31	-4
C	+33	+33	-10	-49

Table 25. Wolof percentages with vs. without C_1=C_2 sequences

Instead of considering the combinations of identical consonants with respect to the whole dictionary, we shall limit ourselves to the expected number of C_1=C_2 words

compared to the total set of consonants made at the same place of articulation. For example, the list of Wolof words which contain a TVT sequence (where T = any dental/alveolar) provides 988 sequences. The frequency of C_1 *l* is 16% among the TVT sequences. Its frequency in C_2 position is 28%. In the absence of mutual influence, we should find 988 x 16% x 28% = 45 occurrences of *lVl* sequences in the Wolof dictionary. It turns out that we find 47. We can therefore conclude that *lVl* either escapes the effects of SPA, or that the effect of SPA is canceled out by the inverse tendency to favor identical consonants, in this case *l*'s. Let us now compare *nVn* sequences: The frequence of C_1 *n* is 14% (among TVT sequences). Its frequency in C_2 position is 18%. We should therefore find 988 x 14% x 18% = 25 occurrences of the sequence *nVn*. In this case we find 41, which represents a deviation of 65% with respect to the norm. This discrepancy is sufficient large to suggest that the sequence *nVn* is relatively privileged among the possible combinations of dental consonants (TVT).

We have chosen these specific illustrations because it is among dental consonants that we find the only negative discrepancies for identical consonants. Within the labial, palatal and velar series combinations of identical consonants are systematically favored. For the dental place, the results are mixed, as seen in Table 26.

ndVnd	nVn	dVd	lVl	tVt	rVr
+132%	+65%	+41%	+4%	-35%	-64%

Table 26. Wolof percentages for dentals, where $C_1 = C_2$

To summarize, two inverse tendencies are at work in Wolof. In a general way, combinations of homorganic consonants are avoided (SPA). However, in the midst of such homorganic consonants, combinations of identical consonants are statistically favored, sometimes reflecting the fact that these sequences are charged with a grammatical function. This second tendency confirms that it is the place of articulation which constitutes the relevant context for statistical biases in the combination of consonants.

But there is something more: we have pointed out several times that our statistical counts show strong tendencies, even with a poor knowledge of the phonology of the language studied. Languages often have additional features which, if taken into consideration, might make the counts more accurate or revealing. Thus, for every language, a good knowledge of the phonology could help find some otherwise hidden tendencies. Let us illustrate this with Wolof again. In Wolof, there is a vowel length contrast that has not been taken into account for the tables presented here (table 9 p. 7 and table 25 p. 26 above). It is interesting, however, to calculate separate tables for short and long vowels respectively. The result is as follows:

C_1VC_2	P	K	T	C
P	– –	–	+ +	+
K		– –	+ +	
T	+	+ +	–	
C	+ +	+	–	– –

C_1VVC_2	P	K	T	C
P				
K	–			+
T	+			
C		+		

Table 27. Wolof percentages with C_1VC_2 vs. C_1VVC_2 sequences[17]

There is an enormous difference between these two tables. For CVVC sequences, there is no trace of SPA! Only four cells show a slight deviation from the expected frequency, but there is nothing systematic here. So, it seems that the "phonological distance" between two consonants is too big for the SPA effect to appear. This kind of analysis has not been made for other languages with vowel length contrast. Therefore, we cannot consider this

[17] The -CVC- and -CVVC- counts are based on 5,991 CVC sequences and 2,104 CVVC sequences respectively.

phenomenon as a universal. But it might well be, for the theoretical explanation involving "phonological distance" seems reasonable enough. In addition to the lack of SPA in the CVVC table above, we observe a reinforcement of SPA in the CVC table, which is quite logical, given that SPA was sensible even on the global corpus.

The problem of reconstructed languages

As pointed out by an anonymous reviewer, the use of reconstructed lexical forms could introduce some bias in the tables. There are several reasons that make us think that we still can use these. First, the nature of the lexicon is different: in a reconstructed one, there are usually no borrowings, or ideophones. And it is precisely those items that can blur the observed tendancies, by having irregular phonological shapes. So the observed tendancies can only be stronger. However, this is not always the case. For example, the Proto-Bantu lexicon as elaborated by the Tervuren group contains all the reconstructed dialectal variants of all zones of Bantu. Thus, the proto-lexicon has far more items than any of the present-day Bantu languages surveyed for this study. Here we can expect the statistical tendancies to be slightly different, and that is the reason why we included not only the Proto-Bantu table, but also four present-day Bantu languages.

Second, concerning diachronic aspects of the problem, it is important to determine if the SPA phenomenon results from historical processes of dissimilation or if its effects are purely synchronic. If we compare the measurements made of the lexicons of proto languages and their descendant living languages, we find that the differences are of exactly the same order as those observed between the individual living lexicons and their average.[18] This means that the data from the proto language better represent the family of the descendant languages than any one of these languages taken by chance. Thus, the Proto-Indo-European lexicon is more representative of Indo-European in general than is, say, the Albanian lexicon. Consequently, in the absence of rigorously established reconstructions, it is justified to use the average calculated on the basis of the living languages of a family, as we do, for example, with the Altantic languages in §7.

7. SPA and the hierarchy of combinations

So far, we have shown that SPA is a reality for every individual language. But a given language may well show some deviations with respect to SPA, especially as far as superclasses are concerned. For example, K-K combination is avoided in all languages, but K-P is overrepresented in Quechua and Kamilaroi and underrepresented in Basque (see Table 24). This might raise some doubts about the existence of superclasses. A more general question is whether there would be a kind of hierarchy with respect to the respective 'rate of avoidance/affinity' of each of the 16 possible combinations. To address these issues, we need a general overview of the values disseminated in the language-individual tables. This table can be obtained in the following way: In Table 28, for the 31 languages examined, we have put the total number of the six possible E/O values ('+', '+ +' etc.) for each of the 16 combinations of P, K, T and C. For example, among the 31 tables presented above, the T-C combination has never shown up as '+ +', but is attested twice as '+', 10 times as '–', 8 times as '– –', 9 times empty (i.e. with an E/O discrepancy of less than 15%), and twice as non-significant ("n.s.") for lack of sufficient T-C occurrences. In the last column we have put the numeric value corresponding to the sum of the plus and minus values in the preceding columns. This total represents the E/O discrepancy proper to each combination: A positive value represents an overrepresented

[18] Compare, for example, Proto-Bantu in Table 14 with "Average Bantu" in Table 29 below.

combination of consonants, while a negative value signals an underrepresented combination. To highlight the wide range obtained in these values, the rows have been arranged with the greatest negative value (-31) at the top and the greatest positive value (28) at the bottom.

	+ +	+	–	– –		n.s.	Total
K-K			5	26			-31
P-P			6	24	1		-30
T-T			17	6	8		-23
C-T		1	14	2	14		-15
T-C		2	9	7	11	2	-14
P-K	1	1	9	6	14		-13
C-C	2	1	6	8	10	4	-11
K-P	1		8	2	19	1	-9
Total	**9**		**155**		**77**	**6**	
K-C	4	10	1	1	13	2	12
P-C	10	8	1		12		17
C-P	7	10			12	2	17
C-K	9	11	1		8	2	19
P-T	4	17			10		21
K-T	7	14			10		21
T-K	13	14			2	2	27
T-P	18	10			3		28
Total	**166**		**4**		**70**	**8**	

Table 28. Summary of the preceding tables

Two facts are immediately visible:

1. The table is divided into two equal parts of eight positive and eight negative rows each. This means that there are as many overrepresented combinations as there are underrepresented. In addition, the zeros and "n.s." are also equally distributed within the upper and lower halves of the table.
2. All of the combinations of consonants from the same class (represented by the gray cells) are in the upper part of the table, indicating that these combinations are globally underrepresented. This is the concrete trace of the SPA phenomenon.

Another important phenomenon can also be noted: All of the combinations within the same superclass (i.e. peripheral K/P vs. medial T/C) are also underrepresented. Thus, the eight combinations which show a negative total are *exactly the combinations of consonants within these superclasses*, whether the consonants are homorganic or not.

Recall that this result is a global one. In an individual language, one or another of these combinations can be overrepresented, as seen in five of the eight rows in the upper half of Table 28. We also observe that the three remaining rows which lack a positive value are all combinations of homorganic consonants (K-K, P-P, and T-T).

The consequence of the preceding observations is that the combinations with a positive total, i.e. those which are globally overrepresented, are all combinations of consonants *belonging to different superclasses*.

Among the 31 languages examined, 17 belong to Niger-Congo: 12 Atlantic languages and five Bantu languages, among which Proto-Bantu. Afro-Asiatic is represented by two Chadic lexicons. All of the other languages are the only representatives of their group. Thus, in order to avoid any bias which might be due to the consideration of related languages, we have recalculated the general table in such a way that each group of

languages is represented by a single language. For the Atlantic languages and Bantu, we have taken the average of the observed individual values shown in the following tables:[19]

Average Atlantic

	P	K	T	C
P	– –	–	+	+
K	–	– –	+	+
T	+ +	+ +	–	–
C	+	+		–

Average Bantu

	P	K	T	C
P	–		+	
K		– –	+	
T	+	+		–
C	+	+	–	+

Table 29: Atlantic and Bantu average values

For the Chadic group, we have eliminated the relatively small Proto-Chadic lexicon which shows too much internal variability.

The 15 groups of languages now each having one set of values are the following: Atlantic, Bantu, Mande, Kwa, Ubangi (Niger-Congo) ; Sara-Bongo-Bagirmi (Nilo-Saharan) ; Chadic (Afro-Asiatic) ; Malagasy (Austronesian) ; Indo-European ; Nostratic ; Mongolian (Altaic); Basque ; Quechua ; Kamilaroi (Australian) ; Port-Moresby Pidgin English. With these changes, Table 30 represents the recomputed values:

	+ +	+	–	– –	norm	n.s.	Total
K-K			1	14			**-15**
P-P			1	14			**-15**
T-T			6	4	5		**-10**
C-C		1	3	4	3	4	**-6**
P-K	1		4	2	8		**-5**
T-C		2	5	2	4	2	**-5**
C-T		1	6		8		**-5**
K-P	1		3	1	9	1	**-3**
Total	**6**		**70**		**37**	**6**	
P-C	4	2	1		8		**5**
K-C	2	3			8	2	**5**
C-K	5	4	1		3	2	**8**
C-P	2	6			5	2	**8**
K-T	1	7			7		**8**
P-T	3	8			4		**11**
T-P	6	6			1	2	**12**
T-K	10	3			2		**13**
Total	**72**		**2**		**38**	**8**	

Table 30. Summary table re-computed

While Table 30 is comparable overall to Table 28, the tendencies are now even more evident:

1. This time, the four combinations of homorganic consonants are at the top of the table, which signifies that they are the most underrepresented (the totals go from –15 to –6). Among these four combinations, only one positive value occurs, viz. C-C in "Average Bantu".

2. Among the combinations forming the second part of the upper half of the table (i.e. non-homorganic combinations from the same superclass), there does not appear to be any neat hierarchy. One can just point out that the combination K-P is the most "normal" of

[19] The average is calculated by dividing the difference of the number of '+' and the number of '–' by the number of languages. For example, the combination C-C in the five Bantu languages presents the value '+ +' three times, the value '+' once, and the value '–' once. The average is therefore : ((3x2) + 1 – 1)/5 = 1.2, which we round to 1. For the combination C-C, "Average Bantu" will thus have the value '+'.

the four, the totals ranging from –5 to –3. The peripheral combinations are favored twice with '+ +' but only in Kamilaroi, with both K-P and P-K. The medial combinations are also favored three times, but only with '+' (T-C in Basque and both T-C and C-T in Kamilaroi).

3. The lower half of the table, which contains the eight combinations of medial + peripheral consonants reveals a surprising internal structure: The first four combinations are exactly and only those which contain C (totals of +5 to +8). The lowest four combinations are exactly and only those which contain T (totals from +8 to +13). Among all these combinations, only P-C (in Nostratic) and C-K (in Kamilaroi) present a single case each of a negative value.

4. The two subgroups in the lower half of the table in turn present an identical internal structure: Combinations in which the medial consonant precedes the peripheral consonant are preferred to the reverse. T-P and T-K have a higher global value than P-T and P-K. Similarly, C-P and C-K have a higher value than P-C and K-C. This preference can be graphically symbolized in the following way:

Combinations of peripheral consonants (P, K) with dentals (T) are never underrepresented. Even more striking, the combinations T-K and T-P are nearly always overrepresented (thirteen of the fifteen linguistic groups examined). Contrary to what we proposed earlier, this means that more is going on than a simple compensation for the underrepresented restricted sequences. If this were the case, we would not expect the '+' and '+ +' values to represent such a hierarchical structure, rather that each of the "compensatory" combinations would have approximately the same values in Table 30. We are therefore forced to conclude that besides the rather spectacular restrictions concerning consonants of the same (super-)class, certain cross-superclass combinations are "favored" by languages. In other terms, beside "bad" words such as *toad* and *bug*, one finds "good" words such as *dog* and *cat*. This conclusion, which goes beyond the objectives of this paper, merits a detailed study of its own. It does, however, raise a further possibility: If there are good words and bad words, are there also good and bad languages? In fact, the very nature of Table 30 allows us to calculate the average values for each combination. By so doing we obtain what could be an "average" language in Table 31, one which conforms exactly to the tendencies shown by all the languages as a whole:

	P	K	T	C
P	– –		+	+
K		– –	+	
T	+	+ +	–	
C	+	+		–

Table 31. An average language

Obviously Table 31 is not that of any of the languages studied.

Finally, Table 30 allows us to establish a hierarchized list of the constraints:

1. pure SPA: adjacent identical classes are prohibited
2. extended SPA: adjacent identical superclasses are disliked

3. combinations involving dentals (T) are preferred over combinations involving palatals (C)
4. The order ‘medial>peripheral’ is preferred over the order ‘peripheral>medial’

While it must be repeated that these constraints do not describe dynamic processes, it is worth noting that they might be responsible for dynamic effects, as when the Bantu language, Tiene, metathesizes CVP-VT and CVK-VT to CVT-VP and CVT-VK (Hyman, 2006).

8. Conclusions and hypotheses

In the course of this study we have reached the following conclusions:

(i) The phenomenon of similar place avoidance (SPA) previously described for Semitic languages, seems to be a linguistic universal, being observed in languages which are both genetically and geographically unrelated.

(ii) Since the effects of SPA are non-categorical and vary slightly from language to language, SPA is best seen as a *statistical tendency*. This tendency can be observed **in spite of** the following factors that may lower its effects :

(a) Reduplication. Reduplication processes are well documented and often operate on morphological grounds. This leads to numerous sequences of identical consonants separated by a vowel. Not only aren’t these sequences forbidden, but they are rather favoured in some grammatical categories (cf. Wolof demonstratives above, but also ideophones, baby talk and other words with expressive or intensive meaning).

(b) Preference for identical *vs* similar consonants. At the phonological level, a close examination of similar-place consonant sequences shows that SPA may not operate equally on all types of same-place consonant sequences. For example, in Russian, the number of initial CVC sequences involving labial consonants is inferior to what is expected, but there are important discrepancies as for the possibilities of combinations within this subset: *pVp-* or *bVb-* sequences are relatively frequent (while their number is still inferior to what is expected), but *bVp-* sequences, for which the two consonants are “dangerously” similar, are strictly limited to a very small number of borrowings : *baptist* (and a few derived forms), *biplan*, *bipol’arnyj* (‘baptist’, ‘biplane’, ‘bipolar’). Furthermore, two of these show a visible prefix *bi-*. This leads us to the following point:

(c) Sequences containing a morphological boundary may be more tolerant with respect to SPA, as illustrated with two of the three *bVp-* examples in Russian above. This is even more evident for *pVb-* sequences where, aside from two borrowings *publika* (and a few related forms) and *pubertal’ny* (‘public’, ‘pubescent’), all the 147 forms among the 97,328 items of A. Zalizniak’s dictionary (Zalizniak 1977) involve the prefix *po-*, so that the sequences are actually *po-b-*, *b-* being the real initial of the stem. As an other example, Larry Hyman (2006 and pers. comm.) points out that in Chichewa, Ciyao, and many other Bantu languages the unproductive verb extensions *-am-* and *-at-* almost fail to occur following CVP and CVT, respectively. On the other hand, productive extensions such as *-i(t)s-* ‘causative’, *-il-* ‘applicative’ and *-an-* ‘reciprocal’ show no such effects.

(d) As illustrated with the *bi-p-* examples in Russian above, borrowings may be less subject to SPA. This may be due to differences in morphological analysis in source and target languages: if a word contains a morphological boundary in its source language, the speakers of the language that has borrowed it have no conscience of this boundary, and thus treat the word as a mono-morphemic one. Or, to put it in a different way, a

word treated as mono-morphemic in a language may have a "disliked" shape because its origin was bi-morphemic and therefore less sensitive to SPA.

(e) Finally, we are responsible for two additional sources of discrepancies. The first one inevitably arises from our lack of competence for morphological segmentation. For a number of languages, we simply took the data without doing any segmentation at all. The second one is our arbitrary classification of phonemic features into four classes, whereas some languages could require more contrasts, e.g. labiovelar consonants, which have been included in the labial class (Bijogo, etc.) and postvelar consonants, which have been placed in the velar class (Wolof, etc.). Moreover, the particular status of some elements may be different from one language to another. The most problematic case is undoubtedly that of *s* (resp. *z*), which we have always included in the palatal class. In languages where there is a contrast between *s* and *sh* (French, English, etc.), /s/ would better fit in the dental class. In some cases, we have computed wordlists that we found on the Internet. These files generally came with no information about the orthographical conventions. For Quechua and Basque, we could assume that the spelling was influenced by Spanish, but we had no such information for Malagasy, Pidgin English or Kamilaroi.

In spite of all these factors blurring the tendency, it is still present not only in each individual language, but also as an average for all the languages.

(iii) Given its universality, it follows that any counter-tendency in a language must be regarded as an anomaly, e.g. the overrepresentation of sequences of palatal consonants in Bantu (Tables 14, 15).

(iv) While Frisch (1996) has hypothesized that constraints on consonant sequences should be proportional to the number of shared phonological features (with $C_1 = C_2$ being a special case), additional counts not presented here reveal that SPA is more sensitive to some feature classes than others. Our measurements show that the dominant effect concerns *place of articulation*, not manner, nasality, or state of the glottis—which may in fact tend to harmonize (Hansson 2001).

(v) We have shown that SPA effects justify grouping the four places of articulation into two *superclasses*: peripheral P,K (*grave, non-coronal*) vs. medial T, C (*acute, coronal*). While affecting both superclasses, it appears that SPA has a stronger effect on peripheral than on medial consonants.

(vi) As suggested by the Classical Mongolian data, it is possible that an elevated level of SPA effects may be compensated by processes of vowel assimilation, especially by back and round vowel harmony. More such languages need to be investigated, however, to test this potential interaction.

(vii) In the course of our investigations we have noted that the statistical biases attributable to SPA are even more robust if the counts are limited to basic lexical items, i.e. a part of the lexicon that includes fewer derived words, borrowings, and elements with an "expressive" value (e.g. ideophones). We have used dictionaries containing as many as 25,000 entries, but also as few as 318 CVC sequences (in the Nostratic wordlist). Although one might *a priori* tend to doubt results based on such a small number of items, SPA effects were found in corpora of all sizes.

(viii) When examined in detail, restrictions due to SPA reveal internal hierarchies. By compiling all measurements of consonant cooccurrence restrictions for our sample of 15 genetic units (that is, languages or proto-languages representative of their genetic family), we have found that this hierarchy involves not only restriction constraints, but also preference ones. The formers concern classes and superclasses and are ordered as follows:

(a). Pure SPA: adjacent identical classes are prohibited
(b). Extended SPA: adjacent identical superclasses are disliked

Within the four classes P, K, T and C, there exists another hierarchy: peripheral classes (P and K) tend to combine less than medial ones.

The preferences may be attributed a different status. In fact, the more we find restrictions, the more we can expect "preferences" to be compensatory. Thus, their distribution is expected to be arbitrary. While this is often the case for individual languages, the distribution of preferences shows more consistence when we consider the summary of all the data, as presented in table 30. The preferences concern combinations of different superclasses, as expected, and are ordered as follows:

(c). Combinations involving dentals (T) are preferred over combinations involving palatals (C).
(d). The centrifugal order (medial > peripheral) is preferred over the centripetal one (peripheral > medial).

So, not only is SPA worth studying, but we are convinced that the study of preferences, which we can label CPA (for 'Centrifugal Place Assymetry') will lead to many important discoveries.

The above conclusions have relevance not only to synchronic phonology, but also to comparative and historical linguistics. Grammont (1895) and Greenberg (1968) have recognized consonant dissimilation as one of the three important factors playing a role in phonetic change, alongside regular sound change and analogical change. If the phenomenon of SPA is universal, and if Language imposes a certain phonetic contour within the limits of the word, questions naturally arise as to how SPA effects come into being and are maintained in the face of the different diachronic pressures to which the shapes of words are subjected. Is SPA a statistical property of Proto-Language that has survived with different nuances in all of the world's languages? While cases of palatalization and labialization are well-known, most processes of sound change affect features other than place: spirantization / affrication, nasalization, voicing / devoicing, aspiration / deaspiration etc. However, there are changes which affect place of articulation, sometimes limited to C_2, as when final **m* and **p* become *n* and *t* in the history of Chinese (Chen 1973). It is not hard to imagine possible, but as far as we know unattested, SPA effects such as the following:[20]

(i) C_2 *m > n, unless C_1 = dental.
(ii) C_2 *m > n only if C_1 = labial
(iii) C_2 > Ø, if it is identical in place to C_1

Such hypothetical changes, however interesting, appear to be a misapplication of the statistics presented here, which should not be taken for what they are not. SPA is not a law, but rather a universal tendency. There is no categorical prohibition against words containing sequences of homorganic consonants, and hence no expectation that sound changes such as the above will ever take place. More reasonable to us might be cases where certain words or combinations of morphemes within words are avoided if they produce violations of SPA. Nevertheless, the highly predictable nature of the tendency suggests that words that violate SPA may be more susceptible to change than those which don't. All such speculations can and, of course, should be tested against further data.

[20] More reasonable to us might be cases where certain words or combinations of morphemes within words are avoided if they produce violations of SPA. Such speculations of course can and should be tested against further data.

References

d'Alton, Paula (1987). *Le palor: esquisse phonologique et grammaticale d'une langue cangin du Sénégal*. Paris: Editions du CNRS.

Bendor-Samuel, John (ed.) (1989). *The Niger-Congo languages*. London: University Press of America.

Berent, Iris & Joseph Shimron (2003). Co-occurrence restrictions on identical consonants in the Hebrew lexicon: are they due to similarity? *Journal of Linguistics*, 39(1).31-55.

Buis, Pierre (1990). *Essai sur la langue manjako de la zone de Bassarel*. Bissau: Instituto Nacional de Estudos e Pesquisas.

Chen, Matthew (1973). Cross-dialectal comparison: A case study and some theoretical considerations. *Journal of Chinese Linguistics* 1.38-63.

Childs, G. Tucker (2000). *A Dictionary of the Kisi Language with an English-Kisi Index*, Köln: Rüdiger Köppe Verlag.

Caron, Bernard (1991). *Le haoussa de l'Ader* (Sprache und Oralität in Afrika, 10). Berlin: Dietrich Reimer.

Chomsky, Noam & Morris Halle (1968). *The sound pattern of English*. New York: Harper & Row.

Clements, George N. and Elizabeth V. Hume (1995). The internal organization of speech sounds. In John A. Goldsmith (ed.) *The Handbook of Phonological Theory*, 245–306. Cambridge, Mass. & Oxford: Blackwell.

Coetzee, Andries, and Joe Pater. 2006. *Lexically Ranked OCP-Place Constraints in Muna*. Ms, University of Michigan and University of Massachusetts, Amherst. Available at http://roa.rutgers.edu/view.php3?id=1219.

Danay Kamis, Mando Makode, Ganda Nikubu, Maurice Tambyo, Namala Ngarassim & Augustin Goytisolo (1986). *Dictionnaire sara-kaba-na-français, Kyabe (Tchad)*. Sarh: Centre d'Etudes Linguistiques-Collège Charles-Lwanga.

Ducos, Gisèle (1971). *Structure du badiaranké de Guinée et du Sénégal (phonologie, syntaxe)*. Paris: SELAF (Bibliotèque de la SELAF, vol 27-28).

Fal, Arame, Rosine Santos & Jean Léonce Doneux (1990). *Dictionnaire wolof-francais suivi d'un index francais-wolof*. Paris: Karthala.

Fleisch, Henri (1961). *Traité de philologie arabe (vol. I)*. Beyrouth, Imprimerie catholique.

Frisch, Stefan A. (1996). *Similarity and frequency in phonology*. Doctoral dissertation, Northwestern University, Evanston, Illinois.

Frisch, Stefan A., Janet Pierrehumbert & Michael Broe (2004). Similarity avoidance and the OCP. *Natural Language and Linguistic Theory* 22.179-228.

Grammont, Maurice (1895). *La dissimilation consonantique dans les langues indo-européennes et dans les langues romanes*. Dijon: Imprimerie Darantière.

Greenberg, Joseph H. (1950). The patterning of Root Morphemes in Semitic. *Word* 6. 161-182.

– (1968). *Anthropological Linguistics: An Introduction*. New York: Random House.

Guthrie, Malcolm (1967-71). *Comparative Bantu: An Introduction to the Comparative Linguistics and Prehistory of the Bantu Languages*, Vols. I-IV. London: Greggs.

Hansson, Gunnar (2001). *Theoretical and typological issues in consonant harmony*. Doctoral dissertation, University of California, Berkeley.

Hyman, Larry M. 2006. *Affixation by Place of Articulation: Rare AND Mysterious*. Paper to the proceedings of the Rara and Rarissima conference, Max Plank Institute for Evolutionary Anthropology, Leipzig, March 29-April 1, 2006.

Illych-Svitych, Vladimir (1971-1984). *Opyt sravnenija nostraticheskih jazykov* ('Essai de comparaison des langues nostratiques'). Moscou, Nauka, 3 vol.

Jakobson, Roman, C. Gunnar M. Fant & Morris Halle (1952). *Preliminaries to speech analysis: The distinctive features and their correlates*. Technical Report 13. Massachusetts: Acoustics Laboratory, MIT.

Jungraithmayr, Herrmann & Dymitr Ibriszimow (1994). *Chadic Lexical Roots*. Vol. 1: Tentative Reconstruction, Grading, Distribution and Comments ; vol. 2: Documentation. Berlin: Dietrich Reimer, coll. Sprache und Oralität in Afrika, 20.

Kawahara, Shigeto, Hajime Ono & Kiyoshi Sudo (2005). Consonant Co-occurrence Restrictions in Yamato Japanese. To appear in *Japanese / Korean Linguistics* 14 (ed. by Timothy Vance). Stanford, CA: CSLI Publications.

Labatut, Roger (s.d.). *Initiation au peul*. Paris: INALCO. (1994 edition).

de Lespinay, Charles (1991). *Langue et parlers baynunk: lexique comparatif. Compte-rendu d'enquêtes et synthèse de lexiques anciens (17e/18e s. - 1988)*. Paris: Centre de Recherches Africaines. 2e édition.

MacEachern, Margaret R. (1999) *Laryngeal Cooccurrence Restrictions*. New York: Garland.

N'Diaye-Corréard, Geneviève (1970). *Etudes fca ou Balanta (dialecte ganja)* (Bibliotèque de la SELAF). Paris: SELAF 17.

Moñino, Yves (1995). *Le Proto-Gbaya: Essai de linguistique comparative historique sur vingt-et-une langues d'Afrique centrale*. Paris, Louvain: Peeters

Nyländer, Gustav Reinhardt (1814). *Grammar and Vocabulary of the Bullom Language*. London: Church Missionary Society by Ellerton and Henderson.

Pater, Joe & Adam Werle (2001). Typology and variation in child consonant harmony. In Caroline Féry, Antony Dubach Green & Ruben van deVijver (eds), *Proceedings of HILP5*, 119-139. Potsdam: University of Potsdam.

Payne, Stephen (1992). Une grammaire pratique avec phonologie et dictionnaire de kwatay (parler du village de Diémbéring, Basse Casamance, Sénégal). *Cahiers de Recherche Linguistique 1*. Dakar: Société Internationale de Linguistique.

Pozdniakov, Konstantin (1991). Perspectives of comparative studies on the Mande and West Atlantic language groups: An approach to the quantitative comparative linguistics. *Mandenkan* 22.39-69.

Pozdniakov, Konstantin & Valentin Vydrine (1987). Reconstruction of the Proto-Manden Phonological Systems. In V. Porkhomovsky (ed.), *Afrikanskoje istoricheskoje jazykoznanije* ("Linguistique africaine comparative"), Moscou: Nauka, pp. 294-356.

Rose, Sharon & Rachel Walker (2004). A typology of consonant agreement as correspondence. *Language* 80.475-531

Sapir J. David (1971). West Atlantic: an inventory of the languages, their noun class systems and consonant alternation, *in* T. Sebeok (ed), *Current Trends in Linguistics 7, 45-98*. Paris: Mouton.

Segerer, Guillaume (2002). *La langue bijogo de Bubaque*. Louvain-Paris: Peeters, coll. Afrique et Langage, 3.

Segerer, Guillaume (1998). *lexique sua*, unpublished ms.

Starostin, Sergei (1998-2005). STARLING Database: http://starling.rinet.ru/.

Tang, Sze-Wing (2000). Identity avoidance and constraint interaction: the case of Cantonese. *Linguistics* 38–1 (2000), 33–61.

Teil-Dautrey, Gisèle, to appear. Et si le proto-bantou était aussi une langue... avec ses contraintes et ses déséquilibres. To appear in *Diachronica*.

Vydrine, Valentin (2004). *Mande reconstructions*, unpublished ms.

Williamson, Kay (2004). *Ijo reconstructions*, unpublished ms.

Williamson, Kay & Roger Blench (2000). Niger-Congo. In Bernd Heine & Derek Nurse (eds), *African languages: an introduction*, 1-42. Cambridge University Press.

Yip, Moira (1998). Identity Avoidance in Phonology and Morphology. In Lapointe, S., D. Brentari & P. Farrell (Eds), *Morphology and its relation to Syntax and Phonology*. Stanford: CSLI Publications.

Zalizniak, Andrei A. (1977). *Grammaticheskiyj Slovar' Russkogo Jazyka* (Grammatical Dictionary of Russian Language). - Moscou: Russkij Jazyk. Electronic version compiled and provided by Sergei Starostin.

Electronic sources - links verified on December 15, 2006

Basque: http://weblandarbaso.miarroba.com/

Mongolian: http://membres.lycos.fr/brunogml/sub/corps.htm

Kamilaroi: Austin, Peter & David Nathan (1998): http://coombs.anu.edu.au/WWWVLPages/AborigPages/LANG/GAMDICT/GAMDICTF.HTM

Malagasy: Lacroix, Jean-Michel: http://www.zomare.com/zomare.html

Quechua: http://members.tripod.com/~jlancey/Peda/Quecfran.htm

Pidgin English: Barhorst, Terry D. & Sylvia O'Dell-Barhorst: http://www.june29.com/HLP/lang/pidgin.html

Bantu languages accessible on the CBOLD site (http://www.cbold.ddl.ish-lyon.cnrs.fr/):

- Bemba: Mann, Michael (1995).
- Kiga-nkore: Taylor (1959).
- Mpongwe: Moudiama, Patrick Daouda (1994).
- Proto-Bantu: Tervuren Bantu Group (1998).
- Swahili: Rugemalira, Josephat (1993).

Appendix :

	PP	PK	PT	PC	KP	KK	KT	KC	TP	TK	TT	TC	CP	CK	CT	CC	Total	Table
*Bantu	557	812	1,635	205	614	470	1,543	215	999	1,164	1,457	69	724	821	803	338	12,426	14
*Chadic	23	65	205	66	69	16	183	52	133	99	123	48	85	47	82	10	1,306	18
*Gbaya	24	33	163	20	33	23	129	18	43	63	90	5	18	32	65	2	761	11
*Indo-European	58	145	586	117	159	71	455	83	339	308	414	145	50	45	92	18	3,085	20
*Ijo	25	38	131	13	19	14	71	7	62	39	35	6	14	15	17	3	509	12
*Mande	2	23	132	22	8	9	101	6	14	24	61	5	14	27	57	6	511	13
*Nostratic	2	14	47	9	25	12	78	34	18	23	13	7	7	7	19	3	318	21
Balanta	21	24	154	43	34	9	94	24	72	63	109	43	58	49	83	24	904	8; 9
Basque	20	69	344	131	39	47	542	163	106	237	550	324	69	84	335	80	3,140	24
Bemba	655	664	1,638	392	424	247	1,004	320	906	857	1,229	309	513	508	596	391	10,653	15
Bijogo	24	78	249	64	48	23	167	31	133	207	196	53	53	64	93	16	1,499	9
Bullom	39	70	157	40	20	20	100	8	46	101	86	12	18	37	61	12	827	9
Hausa	92	145	279	226	249	166	563	287	293	229	352	167	278	148	247	159	3,880	18
Jaad	44	48	226	95	50	11	127	51	84	56	112	35	77	45	109	30	1,200	9
Joola Kwaatay	59	82	299	140	68	50	222	75	170	164	193	113	132	139	185	92	2,183	9
Kamilaroi	36	79	181	47	68	22	161	47	48	38	110	40	15	11	71	6	980	24
Kiga-Nkore	391	742	1,434	672	747	840	2,423	697	1,059	1,657	2,189	1,030	899	922	1,348	894	17,944	15
Kisi	164	213	436	140	98	119	268	59	252	260	226	64	179	223	192	88	2,981	9
Malagasy	82	133	349	88	69	48	182	44	175	178	261	59	61	94	106	15	1,944	17
Manjaku	119	111	515	156	73	73	238	85	300	226	393	130	145	174	325	82	3,145	9
Mongolian	153	2,108	4,000	1,332	1,233	3,399	11,245	2,525	1,920	12,184	6,654	2,715	1,586	6,082	7,567	1,704	66,407	22
Mpongwe	155	223	411	135	198	155	454	142	257	306	262	79	176	258	215	80	3,506	15
Nyun-Buy	94	220	478	183	179	68	703	141	349	566	472	60	271	231	326	87	4,428	9
Palor	47	97	222	100	84	57	311	106	169	231	138	93	92	159	147	63	2,116	9
Fula (clean)	33	7	109	30	23	9	73	34	65	23	66	19	59	29	76	17	672	9; 19
Fula (raw)	58	24	267	48	34	20	186	49	132	54	364	48	108	46	190	23	1,651	19
Pidgin English	126	94	406	111	108	32	184	49	307	137	266	117	88	47	104	39	2,215	24
Quechua	74	181	402	631	160	118	353	737	141	283	236	417	214	379	259	669	5,254	24
Sara Kaba Na	71	105	384	139	208	172	491	262	188	278	296	177	101	177	186	65	3,300	16
Sua	21	18	74	33	9	9	37	7	34	37	71	18	25	32	52	18	495	9
Swahili	75	131	151	86	76	49	134	74	120	97	90	50	115	98	70	65	1,481	15
Wolof	178	326	1,171	397	283	94	938	306	716	722	1,025	454	445	445	734	222	8,456	9

Table 32. Observed number of every combination for each language examined