Statistics and comparative studies: "quantitative diachrony"

Konstantin Pozdniakov*

INTRODUCTION

This paper is prepared as a guide for the users of *Reflex*, presented in the previous chapter of this collection by Guillaume Segerer. I hope that it can be of interest to specialists who wish to apply statistics to comparative linguistic studies, regardless of the target language. In this contribution I develop and systematize approaches already mentioned in [Pozdniakov 1991; 1993]. Through their presentation I demonstrate that the main problem with the use of statistics in linguistic studies lies not in the lack of sophisticated methods, but in the extraction of really useful knowledge from simple statistical data. This article illustrates how qualitative results can be obtained from applying quantitative criteria.

Our approaches are based on three main axioms:

1. In the proto-language lexicon, each element had a certain frequency. Each diachronic change is followed by an automatic change in phonetic frequencies. If *[k] > [k] before [u, o, a] but *[k] > [c] before [i, e], the frequency of [k] diminishes (in the lexicon) while that of [c] grows. Thus:

2. Each divergence of frequencies in related languages is a result of some historical change(s). For every important difference there should be a historical reason, i.e. some change in the proto-language frequencies for one language or for both.

3. Two types of data exist: qualitative and quantitative. The aim is to synchronize the two domains. If our reconstruction does not explain the divergencies in frequencies between genetically related languages, it means that the reconstruction is incomplete. Each divergency of frequencies should give clues to the diachronic reconstruction.

Taking these three statements as a starting-point, we obtain two alternative sets of facts, one qualitative and the other quantitative, instead of only one. The statistical data can be used in both the synchronic and the diachronic domains.

---

* INALCO, IUF, LLACAN. This work is part of the program *Investissements d'Avenir*, overseen by the French National Research Agency, ANR-10-LABX-0083, (Labex EFL).

Use of statistics in diachrony

In diachronic studies, the discussed strategies can be successfully applied to the resolution of the following problems[1]:

1) To highlight archaic fossilized morphemes inside the roots of lexical stems and morpho-phonological processes which are no longer productive (for example, in cases of ancient consonant mutations).
2) To highlight the major directions of regular phonetic correspondences in related languages at the earliest stages of analysis even before the comparison of the lexicons of two related languages (even preceding the compiling of a comparative dictionary).
3) To carry out the internal reconstruction of a given proto-language based on quantitative data from the modern language. The method here is based on the comparison of frequencies of phonemes in each position (in general or in different word classes: in nouns and verbs, for example). This method is very useful for unclassified languages.
4) To perform etymological analyses and verifications of the elements for a comparative etymological dictionary. This method is particularly important for languages in which the genetic links have not been identified.

Use of statistics in the synchronic domain and in typology.

1) Typology of diachronic changes
2) Morpho-phonological processes (phonotactic constraints)
3) Phonological structures.

What kinds of frequencies can be compared? It is possible to compare units of the same feature (1) or search correlations of features (2)?

1) One feature. Minimal language units usually include consonants, vowels, tones and structures.
Inside the four major fields the following can be compared:

- Frequencies in different positions in the same language (for example, initial and final consonants),
- Frequencies in different languages (for example, frequency of CV roots in Wolof and Gola).

2) Correlation of features. Correlation of frequencies inside the four fields or between four different fields (phonotactic constraints), for example:

- Consonants ~ vowels (inside the same syllable),
- C1 ~ C2 in the combinations CVC
- V1 ~ V2 in the combinations VCV

[1] Problems of lexicostatistics and in particular of use of statistics for the creation of genealogical classifications will not be discussed here because they are well known. My personal opinion is partially presented in [Pozdniakov 2014].

- Tone 1 ~ Tone 2 in the structures CVCV
- Tones ~ Vowels
- Structures ~ Tones, etc.

There are 10 possible parameters presented in the following table:

Table 1

	C	**V**	**Tone**	**Structure**
C	x	x	x	x
V		x	x	x
Tone			x	x
Structure				x

It is possible to compare isolated or grouped elements (for example, frequencies of nasal consonants, labial consonants, central vowels or structures with open syllables). Each type of comparison brings in some specific knowledge – this is a source of hypotheses both for diachronic and typological studies. *Reflex* is a unique platform for such calculations. Different results can be obtained in one click. *Reflex* has three features which facilitate its statistical use: the formulas are already integrated (including formulas for observed data and for deviations); specific tools are used to make the grouping of calculated elements easier; the word list can be viewed at any moment. This permits understanding phenomena not included in the statistical analysis proper.
It is impossible to illustrate here all the possible aspects of a statistical comparison. Let us examine just some of them.

1. CONSONANTS

Let's take the requencies of 2 consonants in 2 Atlantic languages:

Table 2: Frequencies of /s/ and /ʃ/ in two Atlantic languages (%)

	Mankanya	Nyun	Atlantic (average)
s-	0	**5,5**	5,1
-s	0	**3,1**	4,1
ʃ-	**3,2**	0	0,6
-ʃ	**4,2**	0	0,5

It is clear that [s] in Nyun corresponds to [ʃ] in Mankanya in two positions. Moreover, the comparison with average frequencies in the Atlantic family shows that the deviation can be found in Mankanya and not in Nyun. So, we can hypothesize that it arose as a result of diachronic change *[s] > [ʃ] in Mankanya.
Let's compare frequencies between two Mande languages:

Table 3

	b	p	v	f	g	k	d	t	z	s
Guro	+	–	+	–	+	–	+	–	+	–
Yaoure	–	+	–	+	–	+	–	+	–	+

Signs '+' and '-' are used to express deviations (positive and negative respectively) from the average frequencies.

You do not have to be an expert in Mande languages to be able to affirm that voiced in Guro correspond to voiceless in Yaoure. Here both regular and systematic correspondences can be observed. It is important to emphasize that the result was obtained without a comparison of the lexicons of the languages.
Yet, the divergencies in frequencies can indicate other phenomena besides regular correspondences.

Table 4: Frequencies of initial p- in three languages

Temne	Ndut	Noon
6%	10%	17%

All three languages are from the Atlantic family, *i.e.* the languages are genetically related, and these frequencies come from a certain frequency of ***p-** in the proto-language (which we do not know). The numbers expressively say that in some language(s) a change has occurred, but we do not know what kind of change it was. Let's introduce a refining parameter: by changing the parameter of calculation in *Reflex*, we can calculate the frequencies of **p-** separately for Nouns and for Verbs.

Table 5

p-	Temne	Ndut	Noon
Nouns (N)	3%	**8% (+)**	**12% (+)**
Verbs (V)	3%	2% (–)	5% (–)
Sum	6%	10%	17%

The situation observed in Temne does not raise any questions: the frequencies of N and V are equal. A different situation can be observed in the closely related Cangin languages, Ndut and Noon. The frequencies are much higher in Nouns than in Verbs. These frequencies can be explained by factors of a morphological nature. In fact, a specialist has no difficulty recognizing a fused noun class prefix **pV-**. Compare 'tongue' (Niger-Congo **dem*): Ndut *Ø-'pɛrɛm*, Noon *Ø-pereem* (Cangin), Jaad *pu-lḛ̀mḛ*, Biafada *bu-deema*, Baïnunk JBK *bu-lèmès*, Manjak *p-reemint*, Mankany *pə-ndɛmənt*, Banjal *fu-re:num*, Gusilay *fu-le:lum*, Nalu *n-lem* / pl. *a-lem*, Limba *fi-liŋ*, etc. Thanks to *Reflex* we can find this fused prefix in other Cangin languages. It also allows us to perform statistically-based morphological reconstructions and to separate original stems in these languages.
We can recognize the presence of a fused prefix **pV-** in the following three ways:

a) by comparing the frequencies in the Nouns and Verbs (internal reconstruction);

b) by comparing the frequencies of **p-** in initial position in different Atlantic languages: in the Cangin languages it is more frequent than in all the other branches;

c) by comparing the frequencies in the initial and final position.

In another Cangin language, Laala, **k-** in initial position shows an extraordinary frequency, 57%, which means that more than half of the entries in the dictionary

start with **k-**. Can it be interpretared as a noun class prefix or as a verbal morpheme (a marker of the infinitive)? Using the button SPLIT in *Reflex* we can see the detailed calculation: 12% of **k-** are in Nouns and 40% are in Verbs. Looking more closely at the Verbs we can see that we are dealing with **ka-**, a marker of infinitive class.
The extremely high frequency of **k-** diminishes the frequencies of all the other consonants. This means that we cannot use this source for the research of regular correspondences by comparing the frequencies in Laala with frequencies in other languages. Before starting the calculations, we need to cancel this prefix from the query. The ideal dictionary for this kind of statistical analysis would be a dictionary of roots, or of lexical stems. It is worth mentioning that we were able to isolate the prefix **ka-** in the dictionary thanks to statistics. If we had compared in Laala the correlation between C and V in the syllables, we would have found that the combination of **k** with **a** in Laala has an extremely high frequency.
Let's take a fragment of frequencies in Joola Banjal in initial position:

Table 6

	Joola banjal	**Average Atlantic**
b-	**21%**	8%
g-	**37%**	6%
f-	**23%**	5%
SUM	81%	19%

Let's compare the frequencies in Banjal with the average frequencies in the Atlantic languages. Eighty percent of the words in Banjal start with one of the three consonants. We do not need to check the dictionary of Banjal in order to affirm that we are dealing with prefixes.
In order to observe the most general tendencies in the distribution of frequencies in each separate language and in the language family as a whole, we need to group the compared elements by place of articulation and by manner of articulation. It is necessary in particular when the word lists are very small and when the statistics of the phonemes is not representative.
Let us group the consonants in series by manner of articulation. The following is the table of frequencies (%) for an "average" Atlantic language.

Table 7

	initial position	final position
Voiceless stops	**31,8 (+)**	17,3 (–)
Voiced stops	**20,3 (+)**	7,9 (–)
Voiceless fricatives	17,7	21,2
Nasal sonorants	13,4 (–)	**30,4 (+)**
Oral sonorants	12,9	15,9
Voiced fricatives	1,8 (+)	0,4 (–)
Voiced prenasalized	1,5 (–)	3,4 (+)
Voiceless prenasalized	0,7 (–)	3,5 (+)

First of all, it is necessary to emphasize that all the positive divergencies of the finals compared to the initials are connected to the 'nasality' feature. We can thus group the nasal consonants as following:

Table 8

	initial position	final position
Oral stops	**52,1 (+)**	25,2 (–)
Oral fricatives and sonorants	32,4	37,5
Nasals	15,5 (–)	**37,3 (+)**

We can affirm that the differences are important:

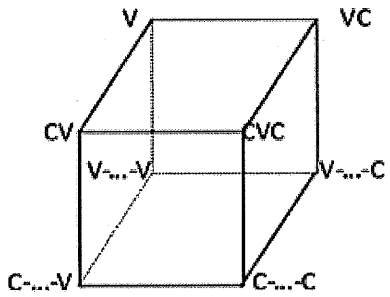
- In initial position half of the consonants are stops while in final position only one fourth are plosives.
- Practically no Atlantic language attests cases where the frequency of the stops in final position is higher than in initial position.
- In the majority of Atlantic languages we can find nasals both in the final and in initial positions (the only language which should be commented on is Bijogo).
- The voiced fricatives are practically absent in the two positions and because of this it was impossible to reconstruct them in the proto-language. For the same reason we cannot reconstruct voiceless prenasalized consonants.

Practically all the series are very stable in the Atlantic languages. The only exception is the series of voiceless fricatives.

2. STRUCTURES

We will focus our attention on 8 essential structures represented in the following cubic model (*scheme 1*):

Scheme 1



The front / back faces of the cube: structures 'Initial C-... / Initial V-...'
The top / bottom faces: structures 'Monosyllabic / Polysyllabic'
The left / right faces: 'Final -V / Final –C'.

The three analyzed examples of the statistical comparison of the structures are illustrated below.

2.1. First dimension (top / lower)

Table 9. Monosyllabic structures opposed to polysyllabic structures (the opposition "top – bottom").

Average Atlantic	Monosyllabic	Polysyllabic	SUM
Nouns	23% (–)	**77% (+)**	100%
Verbs	**47% (+)**	53% (–)	100%

In the majority of the Atlantic languages we examined there are many more monosyllabic Verbs than Nouns. This means that Nouns are longer than Verbs. Why? Only one explication is possible: the presence of fused morphemes in Nouns (markers of noun classes, a phenomenon that we have already found through another procedure.

2.2. Second dimension (front / back)

Let's compare frequencies in 2 structures from Gola – with initial consonants and with initial vowels:

Table 10

	Nouns	Verbs
V-	55,9% (+)	0% (–)
C-	44,1% (–)	**100% (+)**
SUM	100%	100%

It is not necessary to consult the dictionary in order to understand that in Gola Nouns include vocalic noun class prefixes.

2.3. Third dimension (left / right)

Table 11. Structures with final consonants vs. structures with final vowels

%	Basari	Jaad	Gola	%	Basari	Jaad	Gola
Verbs -V	33	13	**97 (+)**	Nouns -V	34	**82 (+)**	**81 (+)**
Verbs -C	**67 (+)**	**87 (+)**	3	Nouns -C	**66 (+)**	18	19
	100%	100%	100%		100%	100%	100%

This table presents very interesting data. The situation in Basari reflects a typical situation for Atlantic languages: the words with final consonants represent the majority of the words in the dictionary (approximately two-thirds of the entries) as the most frequent structure of the lexical stem is CVC. This prototypical situation is valid both for Verbs and Nouns.
In the Nouns of Jaad the opposite situation is encountered: approximately 80% of the Nouns present in the dictionary have final vowels but only 13% of the Verbs. In this language this particular distribution was influenced by two factors. First of all, thanks to this statistical data I paid attention to the fact that in the derivation of Nouns from Verbs (of CVC structure) you have not only a class prefix but also an additional final vowel: *cid* 'to cook' > *ka-cid-e* 'kitchen'; *pees* 'to sweep' > *ka-mpees-a* 'broom'; *puuf* 'to blow' > *ka-mpuuf-ɛ* 'bellows'; *raf* 'to make old' >

ka-ntaf-ɛ 'old age', *etc.* This is also characteristic for other languages of the Tenda-Jaad group. Compare the closely related forms in Konyagi: *i-pas* 'to sweep' > > *æ-fas-a* 'broom' ; bedik *u-wuf* 'to blow' > *gi-mbuf-e* 'bellows'; bedik *u-raf* 'to be old' > *ndaf-a* 'old man'. But in Jaad there was also another cause for the appearance of the final vowel in Nouns. In Jaad there are a lot of loanwords from Mande languages with an initial CVCV structure and the Nouns are much more often borrowed than Verbs[2].

Statistically, with respect to Verbs Gola is different from Basari-Jaad and in Nouns Basari is different from Jaad-Gola. Summing up the data, it is clear that the opposition of Basari to Jaad and Gola Nouns is not of a genetic nature. There are no doubts that Basari and Jaad belong to the same group in the Northern branch of the Atlantic languages. Gola instead apparently does not belong with the Atlantic languages, and represents an independent branch in the Niger-Congo macro-family. This is confirmed by the distribution of frequencies in Verbs: Basari and Jaad are opposed to Gola the only one of the three languages where all the Verbs have a final vowel.

How could such a strange distribution of frequencies occur? It reflects two different independent diachronic changes which took place in Gola, on the one hand, and in Jaad, on the other hand.

In Gola proto-language roots systematically lost the second consonant. Some examples are reported as follows:

- ATL. **jeb* 'cure' (Mankanya *p-jeb,* Nyun *jeb,* ...) ~ **Gola *jwɛɛ***;
- ATL. *ɲamb* 'elephant' (Joola **ɲaab*, Basari *ɲàmb*, ...) ~ **Gola *ó-ɲa̰a̰***;
- ATL. **deng* 'thorn' (Wolof *deg*, ...) ~ Sua *deng-en* ~ **Gola *é-dɛ̰ɛ́***;
- ATL-North. *ɗiŋ* 'cow without horns' (Palor *ɗiŋ*, Sereer *ɗiik*, ...) ~ **Gola *ó-dìì***;
- Balant *tɔg* 'push' ~ **Gola *tɔɔ***;
- ATL. *ɓɔŋ* 'thigh' (Joola **bɔŋ*, Ndut *ɓaŋ*) ~ **Gola o-*gbə̰ə́***;
- Balant *tɔŋ* 'show' ~ Sherbro *ʈonki*, Bom *tɔngi* ~ Nalu *tɔŋ-el* ~ Limba *tɔŋ-ina* ~ **Gola *tɔɔ***;
- ATL. **nof* 'ear' (Bijogo *kɔ-nnɔ*, Cobiana *si-nuf*, Basari *a-nʌ́f*, Palor *nʊf*, Wolof *nɔpp*, Fula **nof-ru*, Baga Mboteni *ɛ-nʌ́f*, Baga Fore *ɪ-nɔ́p*, Nalu *nɛɛw,* ...) ~ Limba *ku-luh-a* ~ **Gola *ké-nʊ́ù***;
- ATL.-BAK **sʊn* 'horn' (Joola **sʊn*, Nalu *seen*) ~ Limba *kɔ-se* ~ **Gola *é-sii***.

This list of examples can be significantly extended. The loss of the final consonant in Gola regularly gives the compensatory length of the vowel. The loss of the final vowels in Gola in the words with the CVC structure can be found by another procedure as well. For example, Table 12 shows a comparison of the frequencies of the monosyllabic words in two different languages:

Table 12

	Gola	Wolof
CV	**38% (+)**	2% (–)
CVC	10% (–)	**41% (+)**
VC	0%	1%
SUM	48%	44%

[2] I would like to thank Guillaume Segerer who drew my attention to this important characteristic of loans influencing frequency distribution.

In both languages the percentage of monosyllabic words is about the same – a little bit less than one half of the lexicon. The two concrete structures have a complementary distribution: apparently the majority of words with CVC structures in Wolof should correspond to words with CV structure in Gola.
To conclude this paragraph it is important to highlight that the comparison of the aforementioned frequencies of eight main structures in Nouns and Verbs suggests some perspectives for the reconstruction of diachronic processes which took place in the proto-language of the family. *Table 13* presents the frequencies of structures in an average Atlantic language.

Table 13

Average (%)	*Verbs*	*Nouns*
CVCVC, CVCVCVC	14 (–)	**25 (+)**
CVCV, CVCVCV	26 (–)	**34 (+)**
CVC	**41 (+)**	19 (–)
CV	5	3
VCVC, VCVCVC	10	11
VCV, VCVCV	4	7
VC	1	1
V	0	0
	100%	100%

The transformation of *CVC in **CV**-CVC in Nouns (integration of noun classes in the roots) is very clear. Despite this, the sum of these two structures in the Verbs (41+14=55%) is much higher than the same sum in the Nouns (19+25=44%). This means that the present explanation is not sufficient. We can postulate the change *CVC > CVC-**V** where the last vowel is a noun class suffix, a determiner or a derivational morpheme.

3. CORRELATION OF VOWELS AND CONSONANTS IN SYLLABLES

Let's group together the vowels in Zaar (West Chadic) with respect to their place of articulation and analyze correlations between consonants and vowels inside syllables: (5353 words)[3].

[3] A capital **U** symbolizes all the back vowels, **A** – all the central vowels, **I** – all the front vowels, **P** – labial consonants, **K** – velars, **T** – dentals, **C** – palatals. For each combination, we compare the theoretical or expected (E) frequency with the actual or observed (O) frequency. If a correlation exists between the qualities of consonants and vowels, this should be manifested by a significant discrepancy between the E and O values. Tables 14-37 present various kinds of E/O discrepancies. The sign '++' means that in the dictionary there is a significant (more than 30%) excess of combinations compared to their expected number (O/E) in accordance with their phonemic frequencies. The sign '+' means a positive deviation of 20% at least. The signs '–' and '-' mean analogical negative deviations in the frequencies of combinations. 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.

Table 14

	U	A	I	SUM
P	++		--	1270
K	+		-	1193
T	--		+	1934
C	-		+	956
SUM	1127	3082	1144	**5353**

Table 14 permits us to show the most general phonotactic constraints in Zaar[4]. They are quite clear:

- Back (rounded) vowels show a strong correlation with labial consonants, a little bit less with velars, which means that back vowels show correlation with peripheral consonants.
- Back vowels combined with central consonants (dental and palatal) are rare.
- Front vowels are often combined with 'central' consonants and do not have combinations with 'peripheral' consonants (especially with labials).
- Central vowels show a total absence of correlations with consonants.

Which result will we obtain if we take into consideration (to simplify the calculation) not all the vowels of Zaar but just the three cardinal vowels: **u, a, i**? It is easy to demonstrate that they are sufficient to highlight all the above-mentioned patterns (*Table 15*).

Table 15

	u	a	i	Tot
P	**++50%**	1%	**--36%**	972
K	+21%	8%	**--36%**	795
T	**-28%**	3%	**++31%**	1489
C	**--36%**	0%	**+29%**	650
SUM	704	2306	896	**3906**

Here the divergencies are presented in percentages compared to the expected numbers.
Firstly, it is evident that the two tables give the same information. This will help us use the same principle of selection for vowels in other languages.
What do the highlighted correlations mean? They very likely demonstrate the diachronic changes in Zaar (if in other Chadic languages they are not found) or in proto-Chadic (if in other Afro-asiatic languages they are not found) or they may have even more common reasons (we cannot exclude the hypothesis of a typological universal). These diachronic changes which cannot be dated using the data from one language, can be related to the change in consonants (for example, ***tu, *cu > pu, ku; *pi, *ki > ti, ci**), or to the change in vowels (for example, ***pi > pu, *ki > ku; *tu > ti, *cu > ci**).

[4] Some of them were formulated in [Boyd 2002] for Chadic languages in general.

In order to estimate whether we are dealing with an innovation in Zaar or whether similar changes can be attributed to a previous period, let's analyze these correlations in two other languages of the same group (West Chadic, South Bauchi).

Table 16

Dott	**u**	**a**	**i**		Guus	**u**	**a**	**i**
P	+		-		P	++		
K	+		--		K			-
T	--	-	++		T	-		
C			-		C			++

This is strong proof that the assimilation of the consonant and of the vowel based on the place of articulation inside the syllable is not an innovation of Zaar. It can be analyzed as a feature of the proto-language. The question of whether we are dealing with a typological universal characteristic remains open.
This hypothesis can be tested on African data. I have chosen 69 languages in the *Reflex* database whose sources are dictionaries with at least 1000 words for each language: 20 Atlantic languages, 10 Bantu languages and various languages from different branches of Niger-Congo. I have also integrated into the analysis Chadic and Nilo-Saharan languages. Now let's look at the correlations of consonants and vowels in our sample:

Table 17

> 15 %	***Positive correlation***			***Negative correlation***			
	u	**a**	**i**	**u**	**a**	**i**	**SUM**
P	**+33**	+14	2	8	3	**-41**	101
K	**+35**	+14	3	8	6	**-51**	117
T	11	3	**+39**	**-30**	-15	6	104
C	8	7	**+39**	**-35**	-10	8	107
SUM	87	38	83	81	34	106	**429**

Table 18

>30 %	***Positive corr.***			***Negative corr.***			
	u	**a**	**i**	**u**	**a**	**i**	**SUM**
P	**+14**	3	1	4	0	**-24**	46
K	**+22**	4	2	4	1	**-31**	64
T	5	0	**+24**	4	1	1	35
C	6	2	**+21**	**-22**	1	2	54
SUM	47	9	48	34	3	58	**199**

The cells demonstrate that a number of languages show an important deviation. *Table 17* shows gaps which exceeds 15% of expected frequencies, while *Table 18* shows those which exceed 30%. Thus, for example, the first cell in *Table 17* indicates that (with a threshold of significant deviation from the expected result equal to 15%) in 33 languages out of 69 a strong positive correlation of labial consonants (symbolized by **P**) and the following vowel **u** inside the syllable.

Forty-one languages out of 69 (with the same threshold of significant deviation) avoid the correlation of labial consonants with the vowel **i**. The most significant correlations are highlighted in bold.

In these condensed tables a lot of interesting information is contained. The most general observation: considering the threshold of deviation of 15% we have 429 deviations from what is expected in 69 languages (6 deviations for each language).

This means that in each language an average of one quarter of the combinations (6 combinations out of 24 possible) deviate, that is ¼ of combinations of consonant and the following vowel are assimilated or dissimilated. Secondly, the sums are more or less equal in different lines, which means that the place of articulation of the consonant correlates more or less identically with a quality of the neighboring vowel. However, in the columns it is not like this: the peripheral vowels (**i, u**) are two times more sensitive to the type of consonant than the central vowel **a**. With the threshold of possible deviation equal to 30% where the possibility of chance is very low, the situation is even more curious. First of all, the dental and palatal (central) consonants (T, C) correlate less than the others (peripherals P, K) with the feature "+back" of the following vowel, and velar consonants (K) are the most sensitive to the neighborhood of back vowels. In 64 languages out of 69, *i.e.* nearly all of them, there is a 30% deviation at least in one combination of a vowel with a velar. These data were most unexpected. In the reconstructions of diachronic phonetic changes, as far as I know, reconstructions of assimilations of consonantal and vocalic segments (as***pi > pu** or ***tu> pu**) are relatively difficult to establish. According to our data this type of change takes place nearly everywhere (our selection included languages from different families and groups). This conclusion should be giver further consideration? If it is confirmed by non-African material, it means that basic statistics gives us the opportunity to look at a new or at least a rarely-used scenario of phonetic changes. In addition, this will allow us to explain the distribution of deviations.

In African languages we can see a systematic correlation of peripheral consonants (labials and velars) with /u/ (also with /a/ but it is less significant) and also of central consonants (dentals and palatals) with /i/. Phonetically clear processes can be traced not only in positive deviations but in the negative ones as well. Negative and positive deviations show complementary distribution. Two principles are pertinent for African languages: to combine the similar elements and to avoid the combinations of opposed elements.

The most frequent phenomenon is the avoidance of combinations of velars and labials with /i/ (51 and 41 languages respectively). There are 106 cases showing restrictions on combinations with /i/. The most neutral correlations are those of the consonants with /a/, also in positive correlations. Languages which do not have significant correlations between consonants and vowels, are a very rare exception in Africa. An example is reported as follows (Table 19).

Other Gbaya languages show deviations typical for African languages.

And finally to conclude the present paragraph, I report a case of language with "atypical" correlations, namely Godie (Table 20).

Unusual in this example is first of all a positive correlation for **Cu** and a negative correlation for **Pu**. I would like to highlight one more time that the selection was limited to cardinal vowels (**u, a, i**) because they are sufficient for the analysis, in most cases. Apparently, it is not always the case. If we group the vowels of this "atypical" language by place of articulation, we obtain the following – completely typical! – distribution (Table 21).

Table 19

Gbaya (Yaayuwee)				9395 records
	u	**a**	**i**	**SUM**
P	-3%	6%	-10%	**2813**
K	7%	3%	-12%	**1939**
T	0%	-5%	10%	**3192**
C	-5%	-4%	12%	**1548**
SUM	**1910**	**5160**	**2422**	**9492**

Table 20

Godié (dadjriwale)				
	u	**a**	**i**	**SUM**
P	--43 %	++40 %	--40 %	**230**
K	+29 %	+25 %	--91 %	**96**
T	-3 %	-11 %	+29 %	**377**
C	++38 %	--34 %	++33 %	**215**
SUM	**245**	**467**	**206**	**918**

Table 21

	u uu o oo ɔ ɔɔ	a aa ɷ ɷɷ ɤ ɤɤ ʊ ʊʊ ɯ ɯɯ	i ii ɪ ɪɪ e ee ɛ ɛɛ	**SUM**
P	-9 %	++41 %	-13 %	**562**
K	++55 %	-1 %	--83 %	**300**
T	-6 %	-13 %	+18 %	**945**
C	-11 %	-24 %	++33 %	**457**
SUM	**1085**	**467**	**712**	**2264**

In this kind of calculation it is very important which scale is selected. If we use the statistical instrument for each single element (for each consonant, for example), we might overlook certain general tendencies. Furthermore, statistics can be less relevant when it comes to detail – especially for rare phonemes. If, on the contrary, we group elements, we might ascribe properties to the system which characterize only some of the isolated elements. *Reflex* permits changing the scale very easily. I personally do prefer to start with grouped elements and then verify the conclusions by using a detailed scale.

4. Correlation of vowels in the sequences VCV

Different languages show correlations between two vowels in the neighboring syllables. Here is a typical example from Wolof.

Table 22[5]

	I	U	A
I	+		-
U		++	-
A		-	+

We can see a diagonal filled with '+' which shows too high frequencies of combinations of the same order in the sequences VCV.

In the following table the Wolof vowels are grouped according to their height:

Table 23

	i,u	e,o,ë,ə	ɛ,ɔ	a
i,u	+	++	-	
e,o,ë,ə		++	--	--
ɛ,ɔ		--	++	-
a,à		--		+

A diagonal is also filled. Saying it differently, the vowels assimilate by order and by series. We will have this distribution if we deal with total assimilation of two vowels. In order to evaluate whether the vowels are identical, let's look at a more detailed table of Wolof.

Table 24

Wolof	**i**	**e**	**ə**	**ɛ**	**a**	**ɔ**	**o**	**u**	**SUM**
i	++	++	--	-			--	+	304
e		++	++	--	--	--	++	++	87
ə		++	++		--	--	--	--	100
ɛ	--	--	--	++		--	--		287
a		--	--		+	-	--	-	1245
ɔ		--	--		--	++	--	-	323
o	++	++	++	--	--	--	++	++	28
u	-	++	-	--		--	++	++	240
SUM	368	50	57	463	972	282	21	358	**2514**

Actually, we can still see a completely positive diagonal: in VCV structures of Wolof identical vowels combine more frequently than expected. At the same time the other positive cells do not seem to be lined up in a chaotic manner – they are concentrated in the four angles of the table. This may mean that in Wolof in the 1st and 2nd degrees there is a tendency to combine equally front and back vowels.

The assimilation of order is also very frequent in other Atlantic languages. Let's see the example from Joola Fogny (Central Atlantic group).

[5] Capital letters signify the grouping of vowels by place of articulation.

Table 25

Joola Fogny	i,ɪ,e,ɛ	a,ə	u,ʊ,o,ɔ
i,ɪ,e,ɛ	+		-
a,ə		+	
u,ʊ,o,ɔ		-	++

We can observe this assimilation in the majority of branches of Niger-Congo. Table 26 shows the distribution we can find in proto-Bantu.

Table 26

*Bantu	u,ʊ,o	a	i,ɪ,e
u,ʊ,o	++	-	-
a	-	++	
i,ɪ,e	-		++

5. CORRELATION OF CONSONANTS IN DIFFERENT POSITIONS

We will look at three types of distant correlations between consonants:

1) Correlation of consonants **by place of articulation** in the CVC sequences
2) Correlation of consonants **by manner of articulation** in the CVC sequences
3) Correlation of **stem initial consonants** with consonants of **noun class markers**.

5.1. Correlation of consonants by place of articulation in the CVC sequences

This phenomena were studied in detail: cf. the paper that was published in [Pozdniakov & Segerer 2007]. We have formulated our hypothesis as *Similar Place Avoidance* (SPA). Later, it was confirmed by a study based on 4500 languages [Mayer et al. 2010]. In the present paper I will limit myself to essential points which can present interest for the methodology of statistical research.
We have shown that in the world languages there is a general tendency of avoidance of combination of consonants of the same order.
Here let's look at three examples demonstrating practically identical distributions of negative deviations.

Table 27

Balant					Basque					Russian				
	P	K	T	C		P	K	T	C		P	K	T	C
P	--	--	++	+	P	--				P	--		+	
K		--		+	K	--	--	+		K		--	+	-
T	+	++	-		T	+	++	-	+	T	++	++	--	+
C	++	++	-	-	C	++			--	C				--

We should bear in mind that each capital consonant (P,T,C,K) indicates grouped characteristics for all the consonants of the main places of articulation. In all the three languages in Table 27 we can see the negative diagonal which confirms our conclusion. This distribution is equally observable in French.

Table 28

French				
	P	K	T	C
P	--		+	
K	++	--		
T	+		-	
C				--

By changing the scale, different details will become more visible. The following Table shows the distribution of frequencies inside oral labials in French:

Table 29

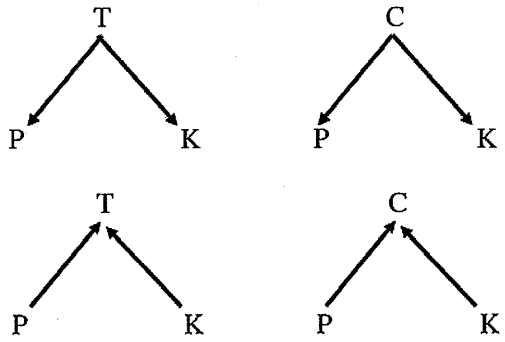
French				
	p	b	f	v
p	++	--		-
b	--	+	++	
f	--	++		-
v			--	++

The situation is typical: French «likes» combinations of identical consonants (for example, **pVp**). It «does not like» to combine minimal pairs in the CVC sequences A dangerous proximity is avoided: too close consonants (**bVp, pVb, f Vp, fVv)** should be assimilated or dissimilated. For this reason there is no word starting with **fVp-**. In Russian there is no such word either, despite a high frequency of these consonants. The combination **bVp-** exists in only three "Russian" words which I can quote without any gloss or translation: [baptist, bi-plan, bi-pol'ar-n-yj]. In English there is just one word with such a combination*:* ***fop*** '1 *obsolete*: a foolish or silly person, 2: a man who is devoted to or vain about his appearance or dress', and no one in French and in Russian.

It is not surprising that this word has an expressive meaning. In these rarely used combinations we can often find ideophones: the particularly expressive meanings call for particularly expressive forms. This is a good example of dialogue between "Nature" and "Culture". The combination **KGB** is extremely expressive in Russian ([kagebe], where **kVg** is very rare), but with the French or the English pronunciation without [g] this combination loses its expressivity.

There is an interesting detail, which is directly related to the problem of interpretation of statistical data. Together with the restrictions (reflected by the negative correlations), languages show also stable preferences for some combinations (positive correlations). Some combinations have a tendency to be privileged in most languages. The most "correct" words according to the statistics can be represented by the following scheme:

Scheme 2



The highest number of /+/ can be found in the following combinations (in decreasing order): **TVP, TVK ; CVP, CVK** (C1 – central, C2 – peripheral) ; **PVT, KVT ; PVC, KVC** (C1 – peripheral, C2 - central). The English words *cat* and *dog* are "good", the word *toad* is "bad".
It is important to underline that we do not deal with compensatory distribution of /+/ due to the presence of /-/. This is confirmed by the fact that in the majority of languages /+/ are systematically found in the same cells.
It was possible to obtain these results only thanks to statistics – the speakers are not aware of these restrictions. For example, they are not aware of the avoidance of the combination **fVp**. Furthermore, these phonotactic constraints are statistical by nature because there are numerous situations in which they are systematically violated. The expressive lexicon was already mentioned. Another important factor which breaks these tendencies of the physiological nature is the presence of a morphological boundary. In Russian again, the combination **pVb** is in general prohibited (only one loan root can be found with this combination, the root **pub-** in *publika, publichnyj,* etc.), but with the prefix **po-** the roots with initial **b-** combine very often: *po-bezhat'* 'start running', etc.

5.2. Correlation of consonants by manner of articulation in the CVC sequences

Besides the correlations of consonants by order, in the majority of languages it is possible to find correlations by series in the sequences.
Table 30 is an illustration from *Joola Fogny*:

Table 30[6]

	P	B	F	W	MP	MB	M
P	+	+	+		-	-	-
B		+					-
F		-	+		++	-	
W					++	++	+
MP	-	--	--	-	--	--	++
MB	--	--	--	+	--	++	++
M		-	-		-		+

A positive diagonal visible in Table 30 means assimilation of the majority of consonants by manner of articulation (except voiceless prenasalized that combine mostly with nasals). In order to define whether they are the same consonants or combinations of different consonants of the same series, a more detailed table should be examined. If we focus on the feature 'voiced' we can see that the same distribution characterizes languages of Niger-Congo, Bantu languages included. Here are some examples that illustrate it (*tables 31-34*):

Table 31

Kru family: Godié	p, t, c, k, kp	b, d, j, g, gb
p, t, c, k, kp	+51%	–68%
b, d, j, g, gb	–64%	+84%

Table 32

Atlantic family: Wolof	p, t, c, k, kp	b, d, j, g, gb
p, t, c, k, kp	+	-
b, d, j, g, gb	-	+

Table 33

Bantu, zone A: Njem	p, t, c, k, kp	b, d, j, g, gb
p, t, c, k, kp	++	-
b, d, j, g, gb	--	+

Table 34

Bantu, zone S: Ndebele	p, t, c, k, kp	b, d, j, g, gb
p, t, c, k, kp	++	-
b, d, j, g, gb	-	+

However in some Bantu languages the reverse situation can be encountered:

Table 35

Bantu, zone F: Nyamwezi	p, t, c, k, kp	b, d, j, g, gb
p, t, c, k, kp	-54%	+81%
b, d, j, g, gb	+21%	-32%

[6] Table 30: P – all voiceless stops, B – all voiced stops, F – all fricatives, W – all oral sonorants, MP – all voiceless prenasalized, MB – all voiced prenasalized, M – all nasal sonorants.

Through this simple procedure, we have just "re-discovered" Dahl's law. It was postulated from the Nyamwezi languages: a voiceless stop becomes voiced when immediately followed by a syllable with another voiceless stop – *e.g.* Nyamwezi **tatu* 'three' > *datu*. It could be the law that we find in one of the Mel languages, namely Temne (Table 36).

Table 36

	P	B	F	W	MP	MB	M
P		++				-	-
B	+	--	-		+		-
F	+		--				
W				--		++	++
MP	--	+		++	--	--	++
MB	--		++	++	--	--	
M	-		++			-	

Contrary to the case of Joola Fogny we can see in Temne a negative diagonal: the consonants of the same series avoid the combinations in the CVC sequences.

5.3. Correlation of stem consonants with the consonants in the noun class markers

The case of Wolof seems to be unique. The results of statistical analysis and internal reconstruction of initial consonants in proto-Wolof show an interesting phenomenon for the theory of comparative studies. In Wolof the noun class markers do not appear on the Nouns, except for fused marks. The noun class markers exist only on the dependent forms, including the definite markers after a noun: *guy G* 'the baobab'. However, the quality of the initial consonant of the lexical base is dependent upon the noun class. For example, Nouns in the noun classes **M** and **L** have a "strong" initial consonant, prenasalized in voiced series (*ndab L* 'calabash') and plosive in voiceless series (*cin L* 'pot', *po M* 'game'). On the other hand, the majority of Nouns in class **W** have a "weak" initial consonant: a sonorant or a voiceless fricative (*fas W* 'horse', *was W* 'carp').

This type of correlation between classes and initial consonants of lexical roots is typical of Northern Atlantic languages. The basis of the consonant alternations is the manner of articulation (plosives, fricatives, prenasalized, etc). The particularity of Wolof is that changes which concern *also* the place of articulation: the initial consonant of the Nouns has a tendency to assimilate by place of articulation with the consonant of the noun class. Thus, in Wolof there is not a single noun of the class **L** which has an initial labial consonant independently of the degree of consonant alternation in this class: there are no Nouns which start with **mb-, p-, m-, f-, w-**. The link between the consonants of noun classes and the initial consonants of stems is represented in Table 37.

Table 37

Wolof	M	W	G	J	L
labials-	+103%	+47%	-32%	-59%	-97%
velars-	-46%	-23%	+99%	-47%	-18%
dentals-	-41%	-16%	-10%	+42%	+67%
palatals-	-35%	-18%	-29%	+57%	+48%

In table 37 the correlation between the place of articulation of initial consonants in the lexical stems (rows) and the place of articulation of consonants in the markers of the noun classes (columns) is illustrated: in the classes with labial consonants (M, W) labial consonants are much more frequent than expected; in classes with velar consonants (**K**) – velars are more frequent than expected; in classes with "central" consonant (**J, L**) – central (dental and palatal) consonants are encountered. The deviations from expected numbers are incredibly high. For instance, the ratio of observed initial labial in the noun class **M** is equal to +103%. This means that in the dictionary [Fal, Santos, and Doneux, 1990] we would expect to find 85 such words, but there are actually 173 of them[7].

CONCLUSION

In this paper we have examined five domains of application for statistics, in addition to the others mentioned in the introduction. All the domains of application can be approached by following the same way of calculation and analysis. The strategic point of this standpoint is that quantitative and qualitative divergences observed in genetically related languages possess the same importance. The statistical results permit a doubling or even a tripling of the number of devices that can be used for diachronic analysis. To conclude, I provide below some suggestions based on my personal experience using statistical data in comparative studies.

1) **Defining the basis of comparison**: For the analysis of the correlation between two elements it is necessary to define the basis for comparison, i.e. the number which you would expect to observe if these correlations did not exist.

2) **Pertinence of the basis of comparison:** If the basis for comparison is limited to one or two units, it is not worth calculating the deviations, especially in percentages. If instead of one expected example you have two or zero, you will find a gap of ±100 percent, but this result would not be pertinent – or even helpful, as it can prevent you from noticing truly interesting phenomena. Quantification of derivations had better be excluded in that case. There are some special methods of calculation to neutralize this factor (such as χ^2), but by applying them you lose the transparency of the results. For significant numbers (starting from around 15 units), the pertinent gap can be defined as ± 30 percent (for 15 units it will be more than 20 or less than 10).

[7] This characteristic of Wolof was studied in detail in [Robert–Pozdniakov 2015].

3) **The grouping of the calculated units:** If the data are not substantial, the calculated units can always be grouped together. It is important to choose the most suitable scale for the comparison: isolated elements or grouped elements, or a selection of some elements, etc.
4) **Degree of precision in the results**: We need to find a reasonable compromise between precision in the calculations and the visibility and transparency of results. For example, you should decide if you really need to fill the table with precise values or with approximations which could be more suitable for your case.
5) **Interpretation of results:** It is important to distinguish between compensatory deviations and pertinent deviations.
6) The most essential principle is to understand what you are looking for before you undertake the calculations.

BIBLIOGRAPHY

Boyd R., 2002, *Bata Phonology: a reappraisal*, Languages of the World 27, LINCOM Europa.
Fal A., Santos R. & Doneux J.L., 1990, *Dictionnaire wolof-français suivi d'un index français-wolof*, Paris, Karthala.
Mayer Th., Rohrdantz Ch., Plank F., Bak P., Butt M. & Keim D. A., 2010, Consonant Co-occurrence in Stems Across Languages: Automatic Analysis and Visualization of a Phonotactic Constraint, in *Workshop on NLP and Linguistics, Finding the Common Ground*, Proceedings of the Workshop. Stroudsburg, PA, Association for Computational Linguistics (ACL), p. 70-78.
Pozdniakov K., 1991, Perspectives of comparative studies on the Mande and West Atlantic: An approach to the quantitative comparative linguistics, *Mandenkan : Bulletin semestriel d'études linguistiques mandé* 22, Paris, p. 39-69.
Pozdniakov K., 1993, *Сравнительная грамматика атлантических языков* [*comparative grammar of the Atlantic languages: noun classes and morphophonology*], Moskva, Nauka.
Pozdniakov K., 2014, О пороге родства и индексе стабильности в базисной лексике при массовом сравнении: атлантические языки [Discussion] [On the threshold of relationship and the "stability index" of basic lexicon in mass comparison: Atlantic languages], *Journal of Language Relationship* 11, p. 187-237.
Pozdniakov K., 2015, Diachronie des classes nominales atlantiques : morphophonologie, morphologie, sémantique, in D. Creissels & K. Pozdniakov (eds), *Les classes nominales dans les langues atlantiques*, Köln, Rüdiger Köppe Verlag, p. 57-102.
Pozdniakov K. & Robert S., 2015, Les classes nominales en wolof : fonctionnalités et singularités d'un système restreint, in D.Creissels & K. Pozdniakov (éds), *Les classes nominales dans les langues atlantiques*, Köln, Rüdiger Köppe Verlag, p. 543-628.
Pozdniakov K. & Segerer G., 2007, Similar Place Avoidance: A Statistical Universal, *Linguistic Typology* 12-2, p. 307-348.