WHAT CAN A DATABASE OF PROTO-LANGUAGES TELL US ABOUT THE LAST 10,000 YEARS OF SOUND CHANGES ?

Egidio Marsico
Dynamique Du Langage, Université Lumière Lyon 2, France

ABSTRACT

This paper describes the construction of a database on reconstructed phonological system and some of the preliminary interpretations. The statistical study of the content of over a hundred proto-phonological systems, does not reveal a significant evolution towards either complexification nor simplification, during the last 10,000 years. This means first that language reached a sufficient level of complexity long ago, and second that sound changes must not be viewed as teological processes, thus placing their motivations and explanations elsewhere. We also offer insights on some theoretical points raised by this study.

1. INTRODUCTION

The aim of this study is to test different theories about complexification and/or simplification of phonological systems in their evolution. Two centuries of research in historical linguistics has led to a large number of reconstructed proto-languages, covering most language families. Though they vary greatly in terms of number of languages concerned and time-depth, it is now possible to take a first look at the content of these proto-phonological systems. In order to do so, we created a database – BDPROTO - based on published material, containing for the moment over a hundred proto-languages. We analyzed the overall organization of those proto-phonological systems (number of segments per systems, consonant/vowel ratio, average number of consonants and vowels, frequency hierarchy of the segments...) as well as in more details (proportion of front vs. central vs. back vowels, rounding, different series of consonants, average number of contrastive places of articulations...) and then compared it with similar statistics for modern languages, based on the UPSID database compiled by Ian Maddieson [4].

Such an enterprise is not straightforward, and although many precautions have been taken while gathering and standardizing the data, we can only very cautiously interprete our results. Far from solely dealing with phonological typology, this research also addresses issues about the reconstruction methodology as well as about the real nature of the phonetico-phonological behavior of languages.

We will first present our sample and then our comparative results between 'synchrony and diachrony', and we'll end by discussing more theoretical points.

2. BDPROTO: THE SAMPLE

In order to be able to make generalizations and not skew our analysis, we need to have a sample representative of the genetic classification of the world's languages. Following Rulhen's classification [6], our sample covers sixteen out of his seventeen families, and it also includes two isolated proto-languages (sumerian, pre basque) and one member of a macro-family (nostratic). Although it may sound paradoxal to use a classification obtained by mass comparison to classify proto-languages reconstructed using - or at least claming to use - the comparative method, we should note here that we do not accept this classification as a whole nor as the best representation of the genetic situation, but simply as a practical mean to organize our sample.

Even if we cover more than 90% of the families, we recognize a disproportion in the representation of each family, some have only one element (Khoisan, Caucasian, Eskimo-Aleut, Na-Dene) where some others can have up to 18 (Austric) or 21 proto-languages (Amerind). However, we can balance this apparent desequilibrium noting that the families with the largest number of representatives are also the largest ones (Niger-Kordofanian 12 proto-languages, Austric 18, Amerind 21), thus turning hyper-representation into a better picture of the actual diversity of these families.

We gathered a total of 113 reconstructed proto-languages but we have only kept 95 of them for the sake of reliability and representativity. (see Table 1.) We have eliminated some doubtful proto-languages (proto Hokan, East-Nostratic, macro-Caucasian) and when a particular branch was developed on three or more levels, we did not take into account the middle nodes in order to have a better trace of the evolution. We also dropped reconstructions that were too close in the family tree (eg. West Saxon, pre Anglian Old English too close to Middle English).

Table 1. Number of languages per family in BDPROTO

Khoisan	**1**
Niger Kordofanian	**12**
Nilo Saharan	**4**
Afro Asiatic	**6**
Caucasian	**1**
Indo Hittite	**6**
Uralic Yukaghir	**7**
Altaic	**3**
Chukchi Kamchatkan	**0**
Eskimo Aleut	**1**
Elamo Dravidian	**2**
Sino Tibetan	**4**
Austric	**18**
Indo Pacific	**2**
Australian	**2**
Na Dene	**1**
Amerind	**21**
Isolated	**2**
Macro family	**1**

Another aspect of representativity is the one of time-depth. We have in our sample, proto-languages going back as far as 6000 or 7000 years (Afroasiatic, Indo-European, Central Khoisan) but we also have some of only a few centuries (Attic, Middle English). This problem of time-depth is an important one, the original idea of this project was to draw several pictures of the phonological systems representing different levels of time-depth, which would have been the best way of detecting particular patterns of evolution. The fact is, that the further you go back, the

less rigorous is the representativity of the sub-sample obtained, and the fewer systems you have, thus making it hard for a statistical study. Our decision has been not to distinguish different levels of time-depth when considering the entire sample, but, when possible, to do so within a family, that might enable us to catch family-specific changes.

We want to make some remarks about the codification. A great deal of the work has been done to standardize the different codifications used by the authors (regarding school or family-specific traditions or simply when thinking that the phonetic interpretation of a symbol went without saying). We decided to use the IPA alphabet to describe our systems. Thus, we made correspondances charts for each of our proto-languages (we would like to thank Ian Maddieson without whom several misinterpretations of symbols would have been made).

3. STATISTICAL ANALYSIS

3.1. On systems

We will present the results and the analysis at the same time.

Before going into the detail of consonantal and vocalic systems, we first give - in table 2. - results about systems in general. Following Maddieson [4], we made the distinction between vowel segment and vowel quality, the former covering every vocalic segment, whereas the latter represents only the vowel distinctions based on height, frontness and rounding. This distinction corresponds to primary vs. secondary features, all vowels having at least a value for the three primary features but the secondary ones being not always present.

Table 2. General statistics on proto-phonological systems

	Seg.	Cons.	Vowels	Qualities	V/C
BDPROTO	30.2	21.3	8.6	5.8	0.47
UPSID	31.0	22.8	8.7	6.3	0.40

Table 2. shows the stability of the overall organization of the systems of the two databases, if anything we can note that the difference in the number of segments in UPSID corresponds to an increase of the number of consonants.

If we consider the average numbers of consonants and vowels, we can use the most frequent segments to draw a picture of the prototypical system, that is:

p,b	t,d		c,ɲ	k,g	ʔ	i	’j’	u
m	n		œ	ŋ		e		o
	s,z	ʃ		x	h		’ʎ’	
	l,r					’ɛ’		’ɥ’
w			j			a		

Although this picture reveals a perfectly organized consonantal system, we looked at the languages containing around the average number (between 20 and 22 consonants) and none of them presented this exact pattern, two were very close (Môn and Katouique with 19 out the 21 consonants) the others having only around 14 of these segments. With a slightly different set (tʃ but no palatal stops, f and no x), Maddieson draw the same conclusion about the effectiveness of such a system (i.e. none of the inventory of UPSID is exactly the same as the one obtained with the most frequent segments). However, as we will see when dealing with the subsets (consonants and vowels), they respect the general organization of segments in terms of the proportion of each categories of consonants. If we take into account the prototypical vowel system and consider only the mean of vowel qualities (vowels with no quotes), it is consistent with the systems with 5 vowels /i e a o u/, which the great majority of them have.

3.2. On vocalic systems

Figure 1. shows that for both databases, the class of systems with 5 qualities is the most frequent and that very few languages have more than 9 vowel qualities in their systems, when the system is too crowded in qualities, they generally prefer to add a secondary contrast (length, nasalization....).

This preference for 5 vowel qualities is reinforced by the fact that the most frequent number of height levels is three.

If we now turn toward the details of the composition of these

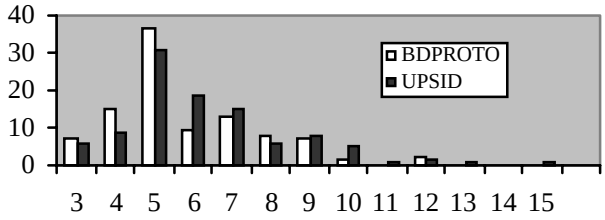

Figure 1. Distribution of vocalic systems according to the number of vowel qualities

systems (see Table 3), once again we can see slight differences between the two stages of languages. If central vowels seems more frequent (and front vowels less frequent) in UPSID, this finds its explanation in the fact that, following the IPA, we describe /a/ as being a front vowel whereas Maddieson choosed to describe it as central. Note the similar percentage of peripheral vowels, revealing their importance in the vocalic organization.

Table 3. Percentage of vowels in different categories averaged across systems

	Percentage			
	Front V. UnR./R	Central V. UnR./R	Back V. UnR./R	Peripheral Vowels
BDPROTO	56 97 / 3	8 94 / 6	37 4 / 96	88
UPSID	40 94 / 6	22 96 / 4	38 6 / 94	88

Table 4. is the most interesting one, it reveals variation in the choice of secondary contrasts, the number of systems using length has almost been divided by 2, and the number of systems with nasalization multiplicated by more than 2.5. Here may lay an illustration of the hypothesis of the cyclic nature of sound changes, a more refined analysis, considering evolution within particular families and taking into account several (if possible) levels of time depth, might show different patterns of evolution for these two contrasts, thus demonstrating the dynamic of systems.

Table 4. Distribution of secondary contrasts

	% systems with contrast	Length	Nasal	ATR
BDPROTO	54	38,8%	8,2%	4,7%
UPSID	45	19,6%	22,4%	

3.3. On consonantal systems

We will only give here results about the places and manners of articulation of consonant.

100% of languages in both databases have at least one plosive, the average number of plosives is 8.5 for BDPROTO and 10.5 for UPSID, distributed in respectively, 4 and 3.6 places, and organized in 2.4 and 1.9 series. The main places are (with almost equal percentages) bilabial, alveolar and velar, the most frequent series being plain voiceless and plain voiced.

95.5% of the systems in BDPROTO have at least one fricative (93.4 in UPSID), the average number being 3.9 (3.9 in UPSID too), divided in 2.5 places in both cases, namely alveolar and glottal (s and h).

If the percentages do not vary a lot with nasals (90% of the systems of BDPROTO for an average number of 3,6, against 97% for UPSID and an average of 2,9), what is interesting is that after a shared prefered place of articulation for alveolar, the pattern is the reverse for the second and the third places. BDPROTO presenting the following order: velar>bilabial and UPSID bilabial>velar. We shall suggest that this might be the result of the temptation to systematically reconstruct a full serie of nasal according to the corresponding plosive serie, and as velar plosive are more frequent, it increases the number of velar nasals in BDPROTO.

Yet, for liquids and approximants we turn back to more ressemblances between the two databases. Both have around two elements of each categories which are present in more than 87% of languages, the favored places being alveolar for liquids (l,r) and palatal and labial-velar for approximant (j and w).

4. THEORETICAL INSIGHTS

4.1. About the evolution of phonological systems

The most remarkable point of this comparative study is the fact that there are no great differences between the "reconstructed" and the "synchronic" levels of phonological systems. The first interpretation is to say that sound changes are not teological and that we cannot talk about an overall complexification or a simplification of phonological systems in the course of their evolution. Of course, relative complexification (and/or simplification) may happen between two particular stages of a language. We said that, due to the quality of our sample, we should be very cautious when interpreting our results, nevertheless interpretations here serve a double purpose: they first are an analysis of the available material, and second they can indicate directions in which to continue the research.

So, what can those directions be ? Saying that sound changes are not teological is not to deny that they happen. They do, they are frequent, important and sometimes they can change a great deal of a system (or a sub-system). What it brings out is that we need to find another motivation (another explanation) to why changes occur. The idea we want to put forward here, is that languages change because they are a "living phenomenon", maybe not intrinsically but definitively in their use by humans.

Everyday, people are communicating, doing their best to respect the articulato-acoustico-perceptual norms of speech, fortunately most of the time with success. But, following John Ohala's approach [5], sometimes they make mistakes when parsing the acoustic signal, by hypo-correction or hyper-correction, thus initiating a potential sound change. Another source of change can be found when considering the acquisition of language by children, although this process may be slower than the preceding one, and not of the same nature if we take into account not only the misparsing errors but the fact that the vocal tract is still in development. Another is the phenomenon of contact, people speaking languages of variable typological distance talk together everyday, sometimes borrowing features (when not segments) from the neighboring community's language and thus modifying their own.

All these external factors force languages to change, but it has long been known, that sound changes are not random but rather directional and, we shall suggest, cyclic, when being driven by physiological constraints. What is interesting then, is to understand the forces (internal factors) that put some limitations on the phonological changes of a particular language. Those can be found in the physical foundations of speech, namely the constraints of the vocal tract, the acoustic and perceptual cues of speech. Also, it appears that, as a complex system, language shows patterns of self-organization, among these emergent properties, we can cite for vocalic systems, the size principle (Lindblom and Maddieson [2]), vowel dispersion (Liljencrants and Lindblom [1]), sufficient contrast (Lindblom [3]) and to some extent focalization (Schwartz & al. [7]).

To summarize this part, we'll say that as no teological purpose is responsible of the constant mutations of languages, we should ascribe it to their living aspect, which, following evolutionary theories dealing with interaction of man with his environment, require languages to adapt to the errors of production, perception, acquisition and to the phenomenon of language contact.

4.2. About the phonological side of languages

What the comparison of proto and modern phonological systems also reveals, is, to some extent, how phonological systems are organized in terms of contrasting elements. Instead of dealing with particular segments, we believe it is advantageous to think in terms of oppositions. The segmentation undertaken by phonology, turning a continuous reality (the speech string) into discrete units, may be a bias against the understanding of how speech works. Far from being singular, both in their production and perception, segments are plural, and the set of variations of their realization is precisely what enables languages to change. Nevertheless, segments are part of a system of opposition where what matters is not really the production of a particular sound itself, but rather the fact that it contrasts with others. It is the general organization of the languages that presents regular patterns, when taken separately, languages reveal a large variation both in synchrony and in diachrony. The idea here is that it is not as important, for example, to find out the exact degree of height of a vowel than to consider how many degrees a language uses.

4.3. About the comparative method

When one reconstructs a proto-phonology, no matter how accurate the phonetic description of the languages compared, one can never be sure of the phonetic interpretation of the proto-segments reconstructed. Not only is the fact that a certain community somewhere back in time spoke the proto-language hypothetical, but the possibility of recovering the entire allophonic system is as well. The point is that this system, we think, is the one that gives the boundaries within which a language can change. Thus, how can we rely on the set of phonemes –rather than sounds- that has been reconstructed, to describe the presence or the absence of such a feature, of such a

contrast, when knowing that we can only reconstruct something that left a trace in the daughter languages ? In other words, it seems like a "natural skew" of the comparative method to only enable to reconstruct what we can access by the existing description of modern languages, thus, consequently showing logical similar phonological systems for both synchronic and diachronic levels.

Nevertheless, we are not saying that comparative work is not worthy of analysis, we just want to draw attention on that in order to point to a potential methodological skew. Moreover, in our opinion, there is no reason to think that languages should have changed drastically during the last ten thousand years, other than within the limits of the natural evolution induced by the use of language, as described in 4.1.

5. CONCLUSION AND PERSPECTIVES

To be sure, these results are preliminaries to further research, though it seems like no great changes happened between the two stages, a refined analysis (almost considering one system at a time) has to be done, and it might reveal evolutions in some smaller domains of phonological systems or also within particular families.

REFERENCES

[1] Liljencrants, J., Lindblom B. 1972. Numerical simulation of vowel quality systems. The role of perceptual contrast. *Language* 48.839-862.

[2] Lindblom, B., Maddieson, I. 1988. Phonetic universals in consonant inventories. In L. M. Hyman & C. N. Li (eds), *Language, speech, and mind: Studies in honor of Victoria A. Fromkin*. London : Routledge. 62-80.

[3] Lindblom, B. 1986. Phonetic universals in vowel systems. In J. J. Ohala & J. J. Jaeger (eds), *Experimental phonology*. Orlando, FL: Academic Press. 13-44.

[4] Maddieson, I. 1986. *Patterns of sounds*. Cambridge University Press.

[5] Ohala, J. 1993b. Sound change as nature's speech perception experiment. *Speech Communication*. 13:155-161.

[6] Ruhlen, M. 1987. *Classification of the world's languages*. Vol. 1: Classification. Stanford University Press. Stanford, California.

[7] Schwartz, J.L., Boë, L.J., Perrier, P., Guérin, B., Escudier, P. 1989. Perceptual contrast and stability in vowel systems: a 3-D simulation study. *Eurospeech 89*. Paris. Vol. 1/2, 63-66.

APPENDIX: LIST OF PROTO-LANGUAGES OF BDPROTO

pre basque
pomo
sangiric
polynesian
lakkja
malayo polynesian
old chinese
eskimo
yokuts
tano congo
quichean
old telugu
dravidian
karen
lower cross
ijo
plang
anatolian
sabaki
môn
new caledonian
philippine
maba
mandekan
katouique
australian
edoid
afroasiatic
cushitic
lower sepik
takanan
early proto finnic
lapp
uto aztecan
mixtecan
popolocan
chiapanec manguean
otomi
zapotecan
chinantecan
zoquean
totonocan
semitic
common berber
lolo burmese
chadic
uralic
samoyed
huon gulf
markham
lakes plain
east tariku
west tariku
ainu
mande
central gur
kimbe
austronesian
ongamo maa
nubian
common nordic
germanic
middle english
indo european
kartvelian
altaic
sumerian
nostratic
central khoisan
mayan
finno-ugric
ob ugric
finno permic
sara bongo baguirmian
gbaya
bantu
gorokan
otomanguean
algonquian
salish
chimakuan
paman
japanese
pre nizaa
guang
attic
vietnamien
coptic
micronesian
maidun
keresan
athapaskan
chibcha
tai