Krajowa Konferencja Zastosowań Matematyki w Biologii i Medycynie

Jastrzębia Góra, 16th–20th September 2013

STATISTICAL ANALYSIS OF PHONEMIC DIVERSITY IN LANGUAGES ACROSS THE WORLD

Stanisław Kacprzak, Mariusz Ziółko, Mariusz Mąsior, Magdalena Igras, and Karolina Ruszkiewicz

AGH University of Science and Technology,
al. Mickiewicza 30, 30-059 Kraków, Poland
{skacprza, ziolko, masior, migras}@agh.edu.pl

ABSTRACT

The results of investigation of the differences among the phonemes of 574 languages all over the world are presented. We attempt to verify the hypothesis of African origin for all languages and gradual languages diversification on other parts of the globe. The obtained results justify the languages classification by applying the methods used in evolutionary genetics.

INTRODUCTION

Language is a complex system of communication. Our ability to communicate in words provides evidence of abstract thinking, constitutes a proof of humanity and distinguishes us from other species. Every spoken language is a system of vocal signs governed by grammatical rules. It allows human beings are able to describe the world, express emotions, verbalize wishes, record history and make plans for the future. Scholars estimate that the number of living languages varies between 6000 and 7000 [1]. The human ability to communicate is both unique and fascinating. Languages evolve and diversify over time. People have always wanted to know how languages had originated and developed. The study of language was initiated by ancient Greeks who tried to dissect its nature [2, 3]. Today linguists focus on the beginnings of language and investigate its history and structure. Languages differ in syntax, morphology and phonology. Many of them have different forms called dialects that are spoken only in certain areas. Languages can be divided into groups. A group of languages that descend from a common ancestor belongs to one language family. This classification helps us to order the majority of languages on a similarity scale but some questions remain unanswered. We still do not know where the cradle of human language is. Atkinson [4] reported a declining trend of phonemic diversity that pointed to the African exodus of modern languages and serial founder model of language expansion. These conclusions are coherent with results presented in [5] that population size of language speakers is correlated with number of phonemes and that smaller populations use smaller phoneme inventories. Atkinson's article provoked severe criticism [6, 7]. One of the objection was raised due to the simplification in measurement of the phoneme inventories in data used for analysis [7]. In this paper we use simple statistical tools to analyse languages but instead of simplified data we use the precise one.

STATISTICAL ANALYSIS OF PHONEMIC DIVERSITY

In order to verify Atkinson's hypotheses we analyse phonemic inventories of languages from 6 continents. Atkinson [4] used data obtained from WALS Database [8], which consists of 504 languages, but as pointed in [7] it suffer from the simplification in measurement of the phoneme inventories. Data obtained by authors in [7] consist of 579 languages, but only 510 were chosen for analysis (some languages were excluded to balance among linguistic families). We analyse 574 languages from data used in [7]. Although we use only roughly 9% of world languages, they cover approximately 50% of world population. Then we diagram the collected data for each continent. The vowel diagrams and the consonant diagrams compare languages with rich phonemic systems with those with poor phonemic systems. Below we present analyses for each continent. We analysed distribution of size of vowel and consonant inventories in languages. We arbitrary choose unimodal distributions (except in one case) that allow us to easily compare them. For every distribution we fit Gaussian function

$$\phi_{\mu,\sigma,\alpha}(x) = \frac{\alpha}{\sigma\sqrt{2\pi}}\, e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \tag{1}$$

using trust–region method, approximated distributions are shown in Figs. (2)–(6).

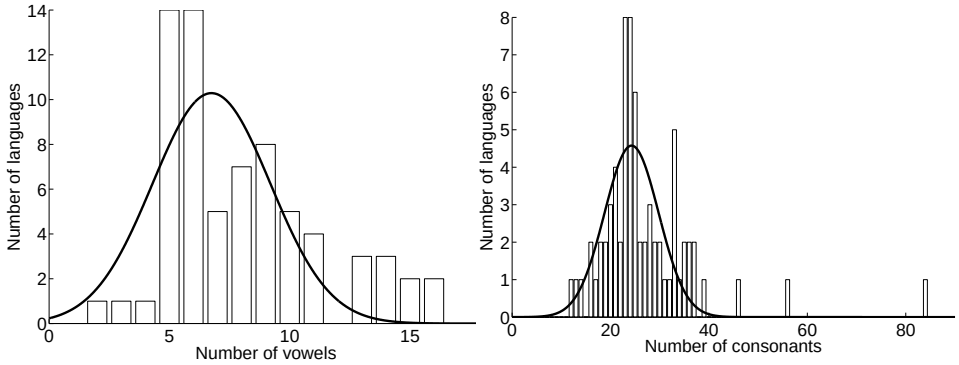

Figure 1. Histograms of vowel (left) and consonant (right) diversity in languages of Europe.

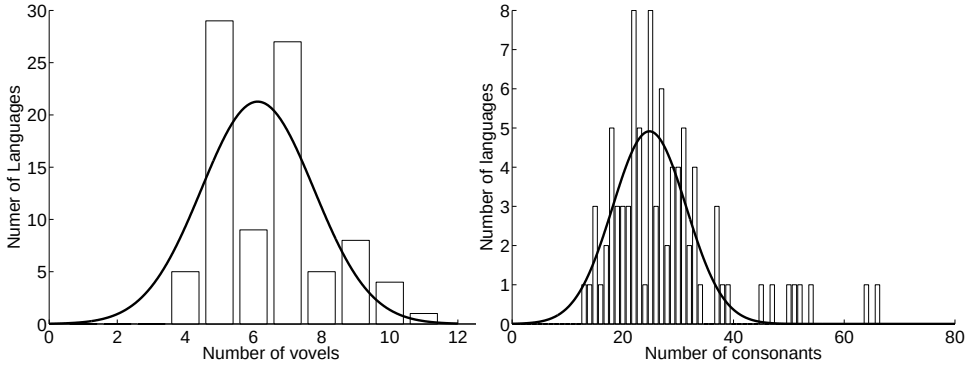

Figure 2. Histograms of vowel (left) and consonant (right) diversity in languages of Africa.

Europe

Our analysis has covered 70 of the numerous languages spoken in Europe (see Fig. 1). The majority of the European population use more than 20 consonants but only a few vowels. Most of

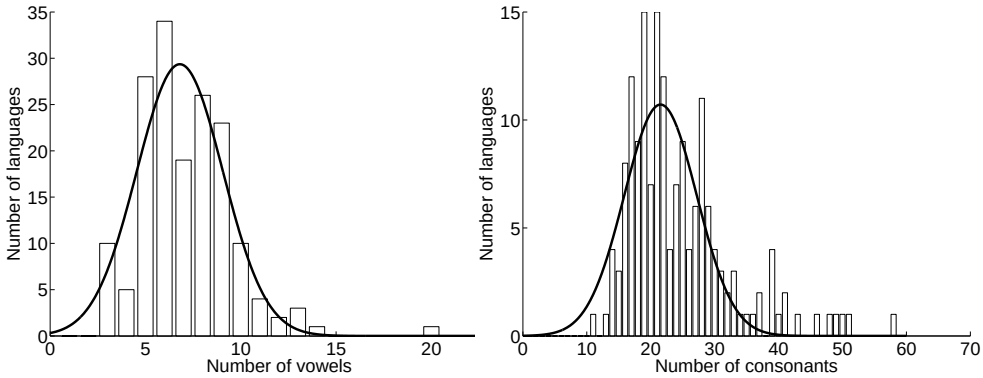

Figure 3. Histograms of vowel (left) and consonant (right) diversity in languages of Asia.

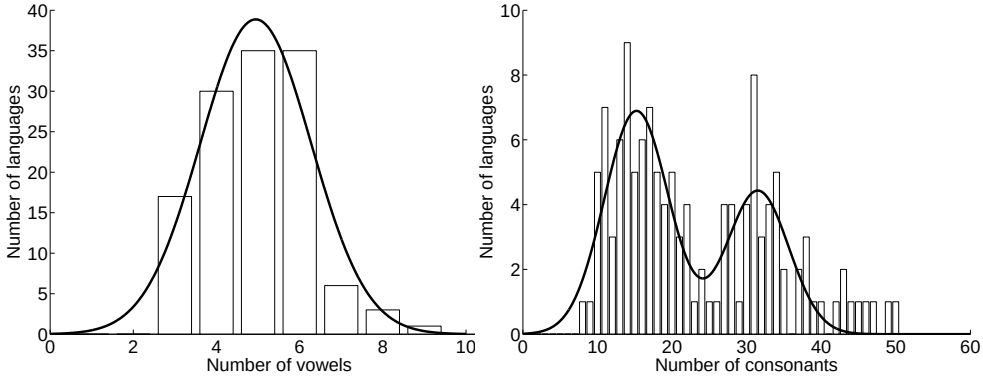

Figure 4. Histograms of vowel (left) and consonant (right) diversity in languages of North America.

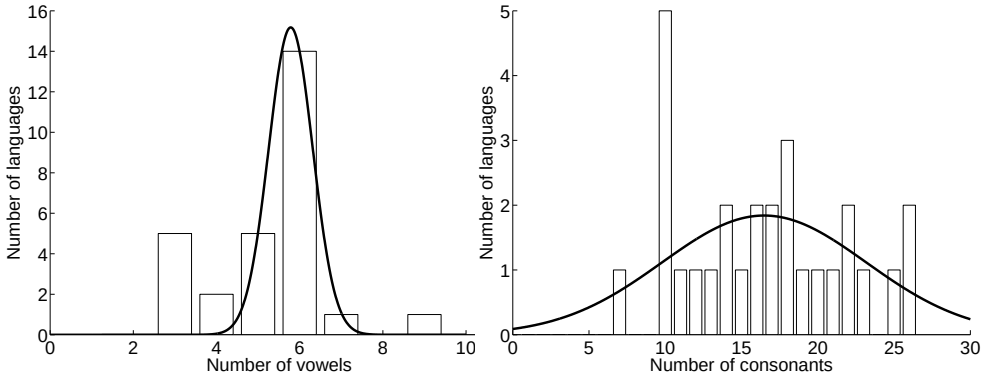

Figure 5. Histograms of vowel (left) and consonant (right) diversity in languages of South America.

these languages have from 5 to 6 vowels and 24 to 25 consonants.The largest group of people use 5 vowels and 36 consonants. There is one language with 84 consonants.

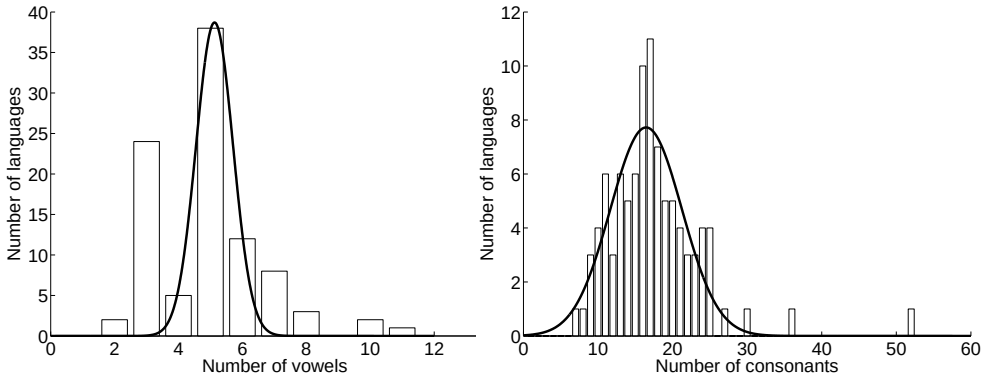

Figure 6. Histograms of vowel (left) and consonant (right) diversity in languages of Oceania.

Africa

According to [4], Africa is the place where the phonemic diversity is the greatest. We have analysed 88 African languages (see Fig. 2). It turned out that in terms of the number of vowels they are comparable to languages from other continents. The inhabitants of Africa use from 4 to 11 vowels. Most of them use 5 vowels. People with rich phonemic systems (11 vowels) are in the minority. What is more, African languages possess between 13 and 66 consonants, but the majority of them have from 15 to 35 consonants. Only 10 African languages have 38 consonants or more.

Asia

We have analysed 166 languages from the Asian continent (see Fig. 3). Asian languages have from 3 to 20 vowels, but the majority of them have between 5 and 9 vowel sounds. Asian languages have up to 58 consonant phonemes. There are 30 languages with 19 or 21 consonants.

North America

The indigenous languages in North America have from 8 to 50 consonants, but the majority of them use between 10 and 20 consonant sounds (see Fig. 4). The vowel systems are not as rich as the consonant ones. The indigenous languages have from 3 to 9 vowels. A major part of the North American Indians use 4 vowels and 20 consonants. Languages with rich phonemic systems (35 to 50 consonants) have relatively few native speakers. Most of American languages have 5 or 6 vowels and 13 consonants.

South America

All the indigenous languages of South America have quite poor phonological systems (see Fig. 5). We have analysed 28 languages and most of them have only 6 vowels and 10 consonants. The majority of the Amerindians use 3 vowels. The languages in Latin America have from 3 to 9 vowels and 8 to 26 consonants.

Oceania

Oceania is the second poorest area (after South America) in terms of the number of phonemes (see Fig. 6). 94 indigenous languages have from 2 to 11 vowels and between 7 and 51 consonants, but there are only 4 languages with more than 26 consonants. So the majority of the languages have between 7 and 25 consonants.

CONCLUSIONS

In Table 1. we show expected number of phonemes in languages on each continent, their standard deviation and number of people that are using examined languages. Table 2. presents parameters of calculated Gauss function (1). Sound diversity among languages differs depending on the continent. All world languages can be divided into four groups. The greatest phonemic diversity can be observed in Africa and Europe. Next is Asia and afterwards North America. Oceania and South

Table 1. Speech sounds statistics (consonant, vowels and all of them) evaluated for languages and for population

		Languages statistics		Population statistics			
	Area	**Expected value**	**Standard deviation**	**Expected value**	**Standard deviation**	**Languages number**	**Population [mln]**
Consonants	Africa	27,74	10,32	28,85	10,63	88	243,2
	Asia	24,48	8,21	25,54	6,54	166	3300,2
	Europe	26,96	10,20	26,32	6,77	70	691,2
	North America	23,42	10,44	23,71	7,96	127	4,2
	Oceania	17,37	6,23	14,57	4,3	95	1,4
	South America	16,43	5,38	24,67	3,21	28	2,3
Vowels	Africa	6,49	1,65	6,25	1,48	88	243,2
	Asia	7,13	2,46	8,63	2,07	166	3300,2
	Europe	8,04	3,27	8,64	3,84	70	691,2
	North America	4,97	1,26	5,07	1,4	127	4,2
	Oceania	4,94	1,71	6	1,89	95	1,4
	South America	5,29	1,41	3,53	1,17	28	2,3
All	Africa	34,23	9,88	35,10	10,23	88	243,2
	Asia	31,60	9,30	34,17	7,32	166	3300,2
	Europe	35,00	9,80	34,96	6,13	70	691,2
	North America	28,39	10,74	28,77	8,98	127	4,2
	Oceania	22,31	6,31	20,57	4,33	95	1,4
	South America	21,71	5,14	28,19	2,24	28	2,3

Table 2. Speech sounds statistics obtained from approximations of distributions with Gauss functions (1)

	Area	μ_1	σ_1	α_1	μ_2	σ_2	α_2	**Root Mean Squared Error**
Consonants	Africa	24,81	6,66	82,04	-	-	-	1,15
	Asia	21,53	5,68	152,57	-	-	-	2,05
	Europe	24,29	5,50	63,13	-	-	-	1,66
	North America	15,23	4,31	74,42	31,51	4,14	45,97	1,39
	Oceania	16,45	4,81	93,10	-	-	-	1,02
	South America	16,48	6,70	30,89	-	-	-	1,03
Vowels	Africa	6,12	1,65	88,07	-	-	-	7,56
	Asia	6,81	2,26	166,06	-	-	-	4,30
	Europe	6,75	2,45	63,25	-	-	-	3,28
	North America	4,94	1,33	129,81	-	-	-	4,42
	Oceania	5,13	0,58	56,28	-	-	-	8,97
	South America	5,79	0,53	20,15	-	-	-	2,23
All	Africa	31,66	6,25	81,70	-	-	-	1,05
	Asia	29,16	6,95	156,82	-	-	-	1,70
	Europe	33,90	7,23	68,45	-	-	-	0,77
	North America	25,75	11,35	129,13	-	-	-	2,02
	Oceania	19,06	2,66	27,50	-	-	-	0,88
	South America	22,43	5,95	30,16	-	-	-	0,84

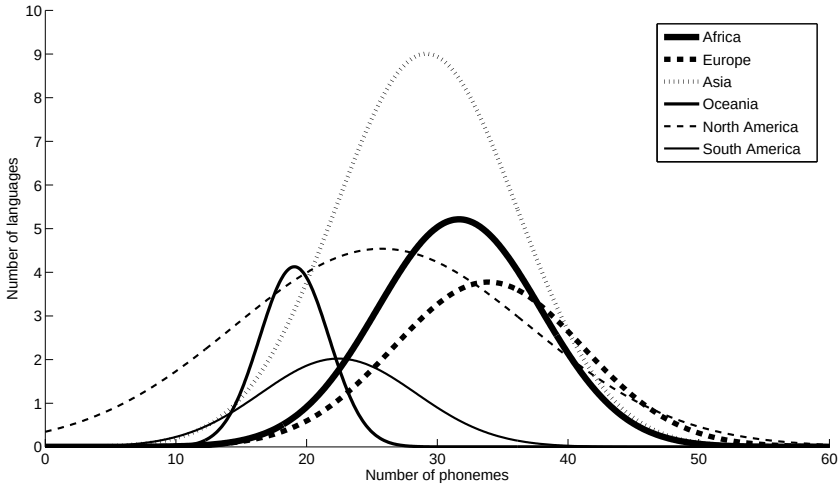

Figure 7. Comparison of approximated distributions Eq. (1) of phonemic diversity across the world.

America belong to the last group marked by relatively low diversity. From Table 1. we see that taking into account population size resulted in increase of expected value of phonemes inventory size, in that sense our results are consistent with [5]. From expected values of vowels inventories (taking into account population size) in Table 1., we see results consistent with [9], where based on analysis of 621 languages five vowel system is predicted as the most original vowel system of human languages.

ACKNOWLEDGEMENTS

This project has been supported by a grant (decision number DEC–2011/03/B/ST7/00442) from the Polish National Science Centre.

REFERENCES

[1] M.P. Lewis (ed.): *Ethnologue: Languages of the world*, 16th Edition, SIL International, Dallas, 2009, Online version: `http://www.ethnologue.com/`.

[2] J. Lyons: *Introduction to theoretical linguistics*, Cambridge University Press, Cambridge, 1968.

[3] R. McKeon: *Aristotle's conception of language and the arts of language*, Classical Philology **41** (1946), 193–206.

[4] Q.D. Atkinson: *Phonemic diversity supports a serial founder effect model of language expansion from africa*, Science **332** (2011), 346–349.

[5] J. Hay and L. Bauer: *Phoneme inventory size and population size*, Language **83** (2007), 388–400.

[6] M. Cysouw, D. Dediu, and S. Moran: *Comment on "phonemic diversity supports a serial founder effect model of language expansion from africa"*, Science **335** (2012), 657.

[7] C.C. Wang, Q.L. Ding, H. Tao, and H. Li: *Comment on "phonemic diversity supports a serial founder effect model of language expansion from africa"*, Science **335** (2012), 657.

[8] M.S. Dryer and M. Haspelmath (eds.): *The world atlas of language structures online*, Max Planck Digital Library, Munich, 2011, http://wals.info/.

[9] L. Hui: *Diversity Analyses on the Basic Vowel Qualities among the Human Languages*, Comm. Contemp. Anthropol. **14** (2008), 42–51.