Scientific Names are Contaminated

Amir Ghazilou[1]*, Afshin Nateghi-Shahrokni[2], Setareh Badri[3]

[1]Shahid Beheshti University, Tehran, Iran; [2] Department of Biology, Malard Islamic Azad University, Tehran, Iran; [3]Payam-noor University, Tehran, Iran

Abstract: The two-part name of a species is commonly known as its Latin name. However, biologists and philologists prefer to use the term "scientific name" rather than "Latin name", because the words used to create these names are not always from Latin**.** The aim of the current study was to evaluate the degrees of similarity between a sample of scientific names and common languages by analyzing relative letter frequency as an onomastic variable. Alphabetical letter frequencies were compared among a sample of scientific names and letter frequencies obtained from nine other languages including: English, French, German, Spanish, Italian, Dutch, Swedish, Latin, and Greek. The dissimilarity between languages was then calculated using Euclidean distance for letter frequencies as a variable. Significant differences were found among different languages and scientific names for all alphabetical letter frequencies examined. Overall, Italian, Spanish and French languages shared highest similarity with scientific names. It can be speculated that scientific naming from Linneaus's first wide-scope goal evolved to the later practice of incorporating names of scientists, with the resulting broadening of the language base from Latin to modern languages.

[Amir Ghazilou, Afshin Nateghi-Shahrokni, Setareh Badri. **Scientific Names are Contaminated**. *Rep Opinion* 2012;4(9):23-25]. (ISSN: 1553-9873). http://www.sciencepub.net/report. 5

Keywords: Latin name; biologist; philologist; alphabetical letter; language

Introduction

The formal system of naming species of living biota is called binomial nomenclature (Schmidt & Bell, 2003). The adoption of a system of binomial nomenclature comes from Swedish botanist and physician Carl von Linne (1707–1778). Linnaeus attempted to describe the entire known natural world, giving every species (plant, or animal) a two-part name (Schmidt & Bell, 2003). This was an improvement over names that involved a sometimes wordy descriptive phrase. The two-part name of a species is commonly known as its Latin name. However, biologists and philologists prefer to use the term "scientific name" rather than "Latin name", because the words used to create these names are not always from Latin, even though words from other languages have been Latinized to make them suitable for this purpose. Species names are often derived from ancient Greek, or from numerous other languages. Frequently, species names are based on a surname, such as a well-regarded scientist, or are a Latinate version of a relevant place name. The aim of the current study was to evaluate the degrees of similarity between a sample of scientific names and common languages by analyzing relative letter frequency as an onomastic variable.

Materials and Methods

Valid scientific name of fishes up to 1 Jan 2011 obtained from fishbase.org and analyzed for alphabetical letter frequency using to Character (Letter) Frequency Count Software 7.0 and an online java applet letter counter at: http://rainbow.arch.scriptmania.com/tools/word_counter.html simultaneously to test accuracy of results. The results were then compared with predetermined letter frequencies obtained from nine other languages including English, French, German, Spanish, Italian, Dutch, Swedish, Latin, and Greek using Kruskal-Wallis one-way analysis of variance and Cohen's kappa coefficient (Zar, 1999). The dissimilarity between languages was then calculated using Euclidean distance for letter frequencies. Cluster analysis was based on the complete-linkage clustering (Zar, 1999).

Results and Discussion

A total of 612864 letters were found to constitute the 32010 valid scientific names with a 0.354 ± 0.67 (mean ± SE) mean occurring frequency. The top twelve most common letters comprised 78.34 percent of the total usage and the top eight letters comprised 58.94 percent of the total usage giving the overall letter frequency sequence as ASIOERUNTLCPMHGYBDVFKXZJWQ, and first letter frequency as PCSAHLBMGETNORDIFKVUXZJYWQ. No definite letter frequency distribution could be detected. In contrast the average sole letter frequency reaches 0.38± 0.64 (mean ± SE) values in English texts (Zim, 1962) and the top twelve letters comprises about 80 percent of the total usage (Zim, 1962) and the top eight letters comprises about 65 percent of the total usage which gives the overall letter frequency sequence as ETAONRISHDLFCMUGYPWBVKXJQZ, and first letter frequency sequence as

TASHWIOBMFCLDPNEGRYUKJVQZX (Zim, 1962). Moreover, the overall letter frequency sequence is evaluated as ESAITNRULODCPMVQFBGHJXYZWK for French (Perec 2001), ENISRATDHULCGMOBWFKZPVJYXQ for German (Beutelspacher, 2005), EAOSRNIDLCTUMPBGVYQHFZJXWK for Spanish (Pratt, 1996), EAIONLRTSCDPUMVGHFBQZJKWXY for Italian (Singh, 2005), ENATIRODSLGVHKMUBPWJZCFXYQ for Dutch (Van Den Broecke, 1985), EIUATSNRMOCLDPQBFGHXYJKVWZ for Latin (McCarty & Cox, 2000), EANRTSLIDOMGKVHFUPBCJYXWZQ for Swedish (Singh, 1999), and AEOINSTUHKRPWMLDGQCFBXZY for Greek texts (Hamer, 2005). Relative frequency of each letter differed among scientific names and tested languages in an ambiguous pattern (Table 1) and when cluster analysis was performed Spanish, French and Italian languages were found to be highly correlated to scientific names in their letter frequency fashion (fig. 1). It can be speculated that evolution of scientific naming, from Linneaus's first wide-scope goal to the later practice of incorporating names of scientists, resulted in broadening of the language base from Latin to modern languages. In a sense, all other languages used as sources for names (toponyms, patronyms) of species could also have contaminated pure Latin. It can be speculated that, coding of the scientific names would face difficulties due to language contamination (Kullbach, 1972).

Table 1. Comparative relative frequency of alphabetical letters among scientific names and eight different languages Note: Asterisk (*) symbol indicates significant difference of relative frequency between scientific names and selected language

	Language									
Letter	Scientific name	English	French	German	Spanish	Italian	Dutch	Latin	Swedish	Greek
A	10.44	8.16	7.63	6.51*	12.53	11.74	7.49	6.80*	9.30	11.13
B	1.92	1.49	0.90	1.89	1.42	0.92*	1.58	1.10*	1.30	0.53*
C	4.89	2.78*	3.26*	3.06*	4.68	4.50	1.24*	3.20*	1.30*	0.77*
D	1.89	4.25*	3.66*	5.08*	5.86*	3.73*	5.93*	2.40*	4.50*	1.87*
E	7.13	12.7*	14.71*	17.4*	13.68*	11.79*	18.91*	9.30*	9.90*	9.87*
F	0.51	2.23*	1.07*	1.66*	0.69	0.95*	0.81	0.80	2.00*	0.59
G	1.98	2.01	0.87*	3.01*	1.01*	1.64	3.40*	0.80*	3.30*	1.63*
H	3.07	6.09*	0.73*	4.76*	0.70*	1.54*	2.38*	0.70*	2.10*	3.96*
I	9.51	6.97*	7.53*	7.55*	6.25*	11.28*	6.50*	8.90*	5.10*	9.53
J	0.19	0.15*	0.54*	0.27*	0.44*	0.00*	1.46*	0.00*	0.70*	0.00*
K	0.41	0.77*	0.04*	1.21*	0.01*	0.00*	2.25*	0.00*	3.20*	3.52*
L	4.97	4.02*	5.45*	3.44*	4.97	6.51*	3.57*	2.50*	5.20*	2.68*
M	3.32	2.41*	2.97*	2.53*	3.15	2.51*	2.21*	4.50*	3.50	2.89*
N	5.99	6.75*	7.09*	9.78*	6.71*	6.88*	10.03*	4.90*	8.80*	8.29*
O	7.87	7.51	5.38*	2.51*	8.68*	9.83*	6.06*	4.40*	4.10*	9.80*
P	3.89	1.93*	3.02*	0.79*	2.51*	3.05*	1.57*	2.20*	1.70*	3.17
Q	0.11	0.09*	1.36*	0.02*	0.88*	0.51*	0.00*	1.40*	0.01*	1.57*
R	6.76	5.99*	6.55	7.00	6.87	6.37*	6.41*	4.90*	8.30*	3.40*
S	10.13	6.32*	7.94*	7.27*	7.98*	4.98*	3.73*	6.00*	6.30*	7.59*
T	5.65	9.05*	7.24*	6.15*	4.63*	5.62	6.79*	6.50*	6.70*	7.54*
U	6.00	2.75*	6.31*	4.35*	3.93*	3.01*	1.99*	8.70*	1.80*	5.83*
V	0.52	0.97*	1.62*	0.67	0.90*	2.10*	2.85*	0.00*	2.40*	0.00*
W	0.17	2.36*	0.11*	1.89*	0.02*	0.00*	1.52*	0.00*	0.03*	3.16*
X	0.4	0.15*	0.38	0.03*	0.22*	0.00*	0.04*	0.30*	0.10*	0.32
Y	1.95	1.97	0.30*	0.04*	0.90*	0.00*	0.03*	0.10*	0.60*	0.14*
Z	0.24	0.07*	0.13*	1.13*	0.52*	0.49*	1.39*	0.00*	0.02*	0.22

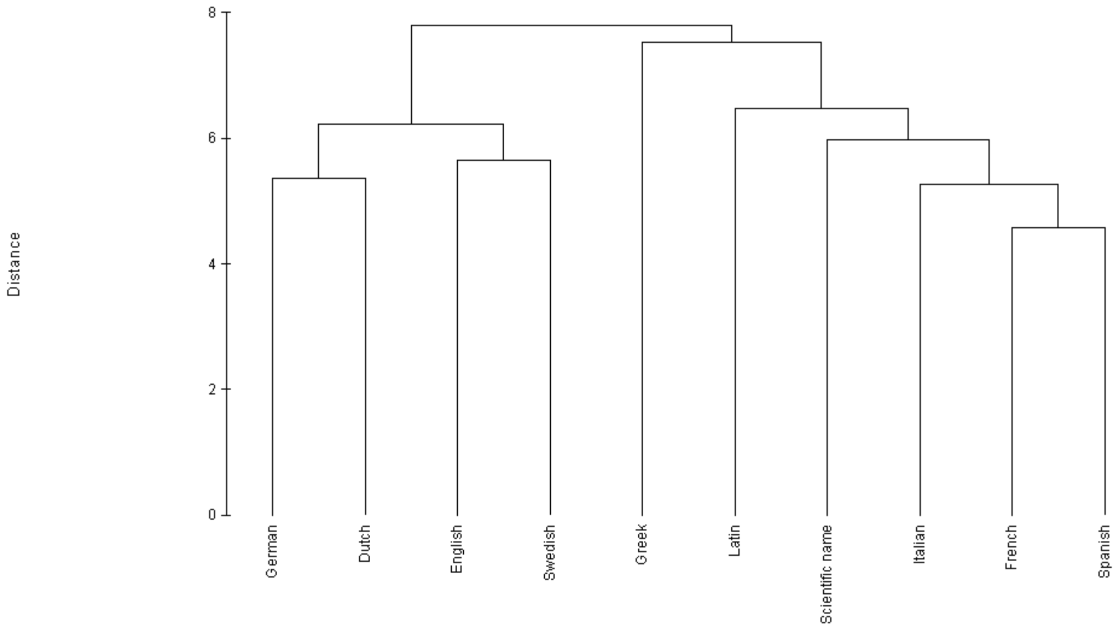

Figure 1. Cluster Analysis of scientific names and different languages by relative letter frequency.

Acknowledgements

Authors are grateful for the funding provided by SRCN organization.

Correspondence to:
Email: A_Ghazilou@sbu.ac.ir
Phone: 00989144041128
Fax: 00984113816094

Submission date: 2012-01-18

References

1. Beutelspacher A. Cryptology. The Mathematical Association of America, New York, Hamer D. Letter frequency statistics. 0052005. http://www.cryptogram.org/cdb/words/frequency.html (accessed: 20 July 2011).
2. Kullbach S. Statistical methods in cryptanalysis. Aegen park press, California. 1972
3. McCarty W. Letter frequency in Latin. Humanist 2000; 14: 314- 320.
4. Perec G. Alphabets. Galilee, Paris. 2001.
5. Pratt F. Secret and Urgent: the Story of Codes and Ciphers. Aegen park press, California. 1996.
6. Schmidt D, Bell G. Guide to reference and information sources in the zoological sciences. Greenwood Publishing Group, Santa Barbara. 2003.
7. Singh S. Code book. Norstedts, Stockholm. 1999.
8. Singh S. Codes and Secrets. Rizzoli, Milan. 2005.
9. Van den broecke M. Ter Sprake: Spraak als betekenisvol geluid in 36 thematische hoodstukken. Foris Publications, Dordrecht. 1985.
10. Zar J.H. Biostatistical analysis, 4th ed. Prentice Hill, California. 1999.
11. Zim H.S. Codes and secret writing. Scholastic Book Services, New York. 1962.

1/18/2012