DOI: 10.2478/v10063-008-0021-8

ANNALES
UNIVERSITATIS MARIAE CURIE-SKŁODOWSKA
LUBLIN – POLONIA

VOL. LXIV, 20 SECTIO AA 2009

Zipf's law in concentration distributions♣

M. Kosmulski
Department of Electrochemistry, Lublin University of Technology,
20-618 Lublin, ul. Nadbystrzycka 38, Poland
Tel./fax +48 81 538 43 55; e-mail: mkosmuls@hektor.umcs.lublin.pl

Concentrations of impurities in fluids often follow Zipf's law, that is, relatively few impurities occur at high concentrations and numerous impurities occur at low concentrations. The concentrations of compounds in air and of elements in ocean water are examples of such distributions. This principle can be used to predict the number of components in a mixture, which occur above certain concentration level, also beyond the range of analytical methods. In most practical applications the existence of minor components can be ignored, but the level of concentration, at which certain component can be ignored depends on the specific problem.

1. INTRODUCTION

The specimens of practical importance, even allegedly pure chemical reagents are mixtures of many components, and actually pure chemical compounds or elements are encountered very seldom. The minor components have usually limited effect on the properties of the specimen, and they are often ignored. Negligence of minor components can be considered as a physical model, that is, a mixture is modelled as a single component (e.g., distilled water as pure H_2O) or as a mixture of a few major components (e.g., air as a mixture of nitrogen and oxygen).

Any extensive property *XYZ* of a mixture can be modelled as

$$XYZ_{\text{effective}}=\Sigma\, h_i XYZ_i \qquad (1)$$

♣This article is dedicated to Professor Roman Leboda on the occasion of his 65th birthday

where

$h_i=n_i$ for $x_i>\varepsilon$ and $h_i=0$ for $x_i\leq\varepsilon$ (2)

where n_i is the number of moles of component i, x_i is its mole fraction, and ε is an arbitrarily selected number <<1.

Likewise, many intensive properties can be modelled as

$XYZ_{effective}=\Sigma\ g_i XYZ_i$ (3)

where

$g_i=x_i$ for $x_i>\varepsilon$ and $g_i=0$ for $x_i\leq\varepsilon$ (4).

This type of modelling is used so often, that the model is often confused with the physical reality, that is, minor components are completely forgotten (treated as non-existing). Indeed, in most practical implications, the negligence of minor components is acceptable, but in several other applications the negligence of minor components leads to erroneous or senseless results. In several equations used in chemistry, e.g., the Nernst equation for the electrode potential a concentration (activity) of a component is in denominator. For such equations, a model, which excludes a possibility of zero concentration of any component is more suitable than a model described by Eq. 1–4. The popularity of homeopathy in the contemporary society is another spectacular example of misunderstanding the state and role of the minor components in solution. The example of homeopathy is discussed in more detail in a separate section.

The main difficulty in the discussions of the minor components is that their concentrations are often below the detection limit of available analytical methods, thus the information about their nature and concentrations are not explicitly available. The model presented below makes it possible to handle a distribution of concentrations of unidentified components. The "Zipfian" approach discussed in the next section is may also be useful in classroom teaching (e.g., of students of medicine) as a tool for semi-quantitative handling of the minor components in fluids.

2. ZIPF'S LAW

Sizes of all sets, which belong to the same category, often have the following distribution:

$$x(r)=Cr^{-\alpha} \quad (5)$$

where x is the size of the r-th largest set, and C and α are empirical constants characteristic for the assembly of sets of interest. Equation (5) expresses Zipf's law, which was originally found for usage of words in English, namely, the most frequent word occurs twice as often as the 2nd most frequent word, 3 times more often than the 3rd most frequent word, etc. Thus, in the original Zipf's law α=1, and Eq. (5) is a generalization. The actual phenomena show deviations from Zipf's law, and the term "Zipfian" is also used for distributions in which $x(r)\approx Cr^{-\alpha}$. Other applications of Zipf's law are: population of cities, family names and other demographic data, income of people, revenue of companies, and other economic data, sizes of files, Internet visits, and other computer-related data, and science citations and other scientometric data [1,2]. Zipf's law is empirical, but numerous explanations for Zipfian behaviour have been offered [1,2]. The hypotheses about the physical basis of Zipfian distribution are not of primary interest in the present study and they will not be discussed here.

In Zipfian distribution the smallest sets are most frequent. In contrast, in many other phenomena a normal or log-normal distribution of sizes of sets is observed, in which the smallest sets are less frequent than middle-sized sets.

I argue that Zipfian approach is useful in description of distribution of concentrations of impurities in actual specimens of fluids. The important difference between the discussed above examples of Zipfian behaviour and the concentrations of impurities is that in glottometrics, economy, demography etc., complete experimental data, also for the smallest sets in the assembly of interest is available. In contrast, the availability of concentrations of the least abundant components in a mixture is limited by the detection limits of analytical methods. In order to test the above hypothesis two well documented data sets were selected. The compositions of other mixtures are usually less well documented.

2.1. CASE STUDIES

2.1.1. Air

Air has nearly constant composition except for the fraction of water, which is variable. Figure 1 shows the volume fractions of components of dry and wet air (chemical compounds) ranked from the most abundant to the least abundant. The results are plotted in log-log coordinates, in which Eq. (5) produces a straight line with a slope of -α. The composition of dry air represented by data points was taken from [3], except the volume fraction of CO_2 was set to 3.5×10^{-4}. Similar composition is reported in other sources [4]. In wet air, a typical value of the volume fraction of water of 0.01 was arbitrarily assumed. For both dry and wet air the composition was normalized to produce a sum of volume fractions of 1. The lines are the best-fit straight lines, with C=0.97 and α=7.07 (wet air), and C=0.68 and α=7.22 (dry air). Four least abundant components were rejected in the calculations. The straight lines corresponding to Eq. (5) reasonably reproduce the actual volume fractions of the 2^{nd}-13^{th} most abundant components of dry air and of the 2^{nd}, 3^{rd}, and 5^{th}–14^{th} most abundant components of wet air, and the volume fractions of the most abundant component (nitrogen), and of the least abundant components are severely overestimated (log scale!). The range of the validity of Zipf's law may be wider than Figure 1 suggests. Namely, the data on volume fractions of components found in literature and used in Figure 1 is not complete. Sum of NO and NO_2 volume fractions appears as one entry, and ozone, and numerous anthropogenic impurities, e.g., freons, chlorine, benzene, and other low molecular organic compounds are neglected. Volume fractions of these components may be higher than the lowest volume fractions shown in Figure 1. The extrapolation of data reported for the 2^{nd}–13^{th} most abundant component by means of Eq. (5) predicts presence of 16 substances at a volume fraction $>10^{-8}$, of 23 substances at a volume fraction $>10^{-9}$, of 32 substances at a volume fraction $>10^{-10}$ (volume fractions) in dry air, which is possible, considering the mentioned above anthropogenic impurities. With exceptions of hydrogen and He, all components considered in Figure 1 have molecular masses of the same order of magnitude, and re-analysis of air composition in terms of mass fractions produces similar results.

2.1.2. Ocean water

Figure 2 shows the mass fractions of elements in ocean water ranked from the most abundant to the least abundant. Almost all natural elements are taken into account in Figure 2. The results are plotted in log-log coordinates. The composition of ocean water represented by data points was taken from [4], except the mass fraction of nitrogen was set to 1.55×10^{-5} (rather than the

originally reported 5×10^{-7}), and the mass fractions were normalized to produce a sum of 1.

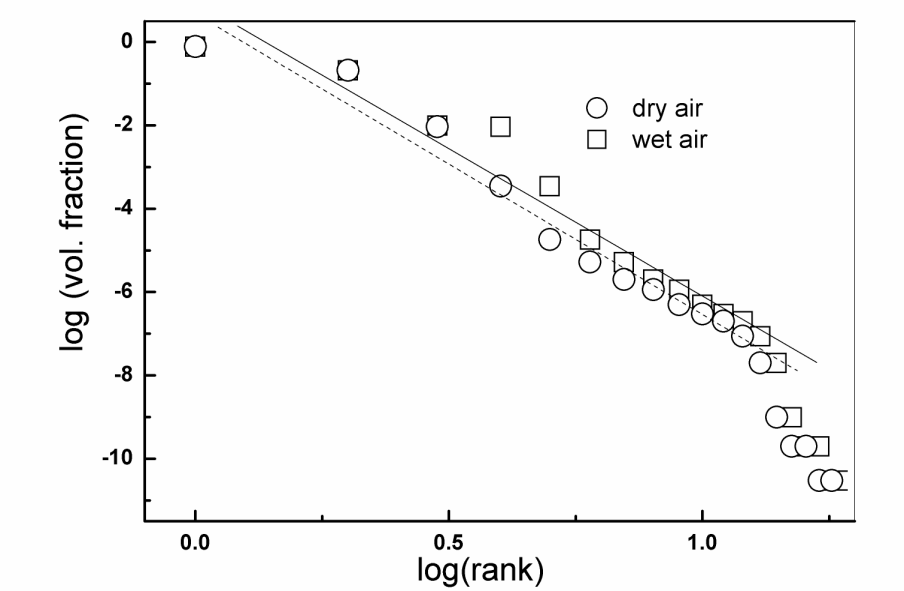

Fig. 1. Distribution of the concentrations of chemical compounds in air.

Mass fraction of nitrogen of $>1\times10^{-5}$ is reported in several sources, which is substantially higher than that reported in [4]. The mass fractions of elements other than nitrogen reported in different sources are consistent. The line in Figure 2 corresponds to Eq. (5), with C=2.7 and α=7.95, and it reasonably reproduces the actual mass fractions of the almost all elements. Only the mass fractions of 2 most abundant elements (oxygen and hydrogen), and of 4 least abundant elements are severely overestimated. Thus the distribution of elements in sea water is nearly Zipfian. The data points in Figure 2 fit Eq. (5) surprisingly well. Usually Zipf's law holds for large assemblies (thousands of words of English, income of millions of people), and fails for small assemblies (aminoacids in natural proteins, letters of alphabet). It should be emphasized that the distribution of elements in Earth crust [4] follows the log-normal distribution rather than Eq. (5).

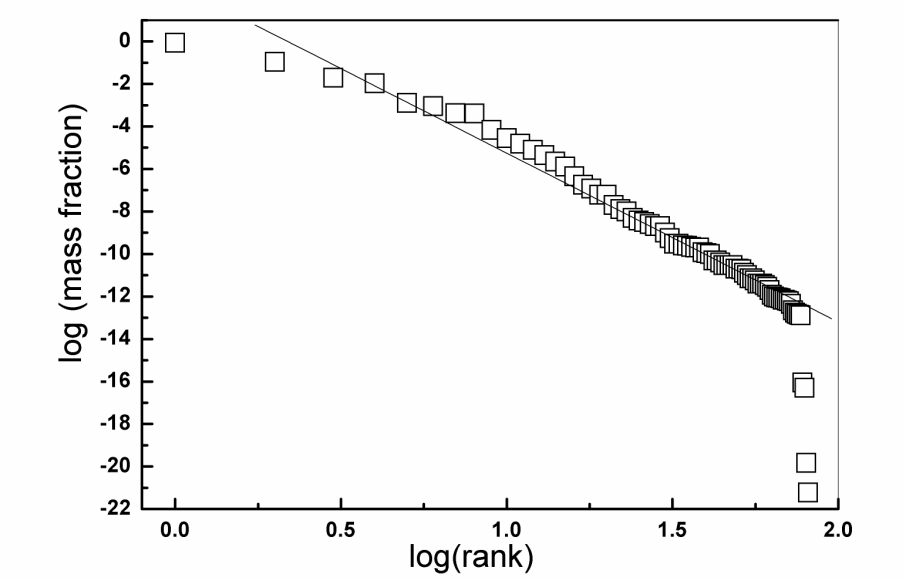

Fig. 2. Distribution of the concentrations of chemical elements in ocean water.

3. DISCUSSION

Air (section 2.1) and sea water (section 2.2) are examples of homogeneous, fluid, one phase-systems, and they can be considered as samples of N_2+O_2 mixture and of H_2O, respectively, both containing impurities (all other compounds/elements indicated in Figures 1 and 2). In both cases distribution of concentration of the impurities is nearly Zipfian. The Zipf's law holds for concentrations expressed in terms of chemical elements or of chemical compounds, as volume (mole) fraction or as mass fraction. Entropy is responsible for occurrence of small amounts of many compounds in fluid phases. It can be hypothesized that in other, less well documented fluid one phase systems, e.g., distilled water, high-purity chemical reagents, etc., the distributions of concentrations of the impurities are also Zipfian. Certainly, the C and α in Eq. (5) in other mixtures can be very different from those found for air (section 2.1.1) and for sea water (section 2.1.2). The parameters of the distribution can be estimated from the concentrations of the most abundant impurities as it is shown in Figure 1 and 2. In two studied cases $\alpha >> 1$, and this is why a few most abundant components constitute almost entire mass (volume, number of atoms) of the sample. With $\alpha = 1$ (observed in other applications of Zipf's law) the contributions of less abundant components to the entire size of assembly is more significant.

Usually the knowledge of concentrations of less abundant impurities is of limited practical importance, and the properties of the mixture depend on the most abundant components. For example artificial sea water for laboratory tests is prepared by dissolution of a few salts in water and most elements indicated in Figure 2 are not used. Very often calculations with very few (e.g., 2 in the case of air) most abundant components are sufficiently precise for certain purposes, and the other components can be neglected. Such an approach is a model rather than physical reality. The real world is Zipfian with $\alpha>>1$, and its model analysis is similar to truncation of terms in a series in mathematics. The number of terms taken into calculation is a compromise between precision and difficulty. The impurities are especially important in adsorption and catalysis, when a minor (in terms of mass fraction) component can substantially affect interfacial tension and reaction rate. Surface tension is often used to test the purity of water. Thus the answer to the question, which components of a mixture are important, and which may be truncated is not always obvious. The surface active components may be difficult to detect by usual analytical methods, and Zipf's law may be useful in prediction of their concentration levels.

4. CONSECUTIVE DILUTIONS

The possibility to obtain very dilute solution by a series of consecutive dilutions of more concentrated solution is an example of practical problem related to the Zipfian approach to concentrations of impurities. For example 1 cm^3 of 1 M aqueous solution diluted with water to 1 dm^3 produces 10^{-3} M solution. The 10^{-3} M solution can be used to obtain 10^{-6} M solution in an analogous way. The next dilution would produce 10^{-9} M solution, etc. However, the number of consecutive dilutions, which still produce solutions of controlled concentration of the solute of interest is limited by the fact, that the solvent (water) used for dilution contains certain amount of that solute, at a concentration, which is unknown, but greater than zero. This concentration is negligibly low by preparation of 10^{-3} M solution by dilution of 1 M solution, but it becomes significant at certain level of dilution, e.g., from 10^{-9} to 10^{-12} M. The lower concentration limit of the method of consecutive solution is determined by the concentration of that solute in the solvent, which varies from one solute to another, and equation (5) (with coefficients estimated from the analysis of solutes, which are within the detection limit of standard analytical methods) allows semi-quantitative approach to this problem. Several examples of very dilute solutions (down to a few atoms per cm^3) allegedly obtained by a method of consecutive dilutions can be found in the scientific literature, and homeopathy is an example of ignoring the Zipfian character of concentration distributions [5]. The principle of homeopathy is based on the assumption that very dilute

solutions (down to a fraction of one molecule per liter) can be obtained by a series of consecutive dilutions, and this assumption is incorrect. Widespread popularity of homeopathy indicates, that the presence of impurities in water at low, but non-zero concentrations is not realized in the community of physicians.

5. SPECIATION IN SOLUTION

Speciation in solution is another problem related to Zipf's law. Let us consider a solution containing a metal cation Me^{2+} and a ligand L^-. These ions form several complexes (species) $[Me_xL_y]^{2x-y}$, which differ in their stability and in their abundance. The number of various complexes may be substantial, but this is not practical to consider them all, for the same reason as it is not practical to consider all possible impurities (*vide infra*). Therefore only a few most abundant (most stable) complexes are taken into account (for example: $[MeL]^+$, $[MeL_2]^0$ and $[MeL_3]^-$), and the other complexes are ignored. The fact that certain complex was ignored does not imply its nonexistence, but it only indicates its negligible concentration. The difference in stability and concentration between the least abundant considered species and the most abundant neglected species may be substantial, but the stabilities (concentrations) may also be uniformly distributed over the log scale (Zipfian distribution), and in the later case, the difference in stability and concentration between the least abundant considered species and the most abundant neglected species may be rather insignificant. The decision, which species should be taken into account, and which should be ignored is a question of subjective choice. For instance in the discussed above Me^{2+}-L^- system, certain authors may consider only 3 complexes: $[MeL]^+$, $[MeL_2]^0$ and $[MeL_3]^-$, but other authors may also consider an additional species, e.g., $[MeL_4]^{2-}$ which is less abundant than $[MeL]^+$, $[MeL_2]^0$ or $[MeL_3]^-$, and more abundant than any other complex. Both models (with 3 and with 4 complexes) involves the stability constants of the complex species. The numerical values of the stability constants of the $[MeL]^+$ $^-$, $[MeL_2]^0$ and $[MeL_3]^-$ complexes (which are considered in the both models) in a model with 3 complexes (neglecting the $[MeL_4]^{2-}$ complex) is different from the stability constant of the same complexes in a model with 4 complexes, and the difference depends on the α parameter in Eq. (5). In other words, a stability constant taken from the literature should be considered as a part of a model involving certain complex species and neglecting other complex species, and it must not be used outside that model. Especially it must not be combined with stability constants of other complexes determined within other models. This statement may seem trivial, but a principle of not-combining stability constants of complexes determined within different models is not generally observed. Examples of combining stability constants of metal complexes taken from different literature sources (and most likely obtained in

the original literature within different models) can be found in several recent publications [6-10].

6. CONCLUSION

The common property of concentration distributions in fluids is that relatively few compounds/elements/species are present at high concentrations, and many compounds/elements/species are present at low concentrations. In most practical applications the existence of minor components can be ignored, but the level of concentration, at which certain component can be ignored depends on a specific problem of interest. The description of the real systems neglecting the minor components is a model, which has certain limitations.

7. REFERENCES

[1] W. Li, *Glottometric*, 5, 14-21. (2002).
[2] W. Reed J., B. D. Hughes, *Phys. Rev. ,* 66, 067103 (2002).
[3] L. Kolditz, *Anorganikum*, 13th edition (in German) Wiley-VCH: Weinheim, p.518. (1999).
[4] *CRC Handbook of Chemistry and Physics,* 79th edition CRC: Boca Raton, pp.14-14, 14-16 (1998-9).
[5] M. Kosmulski, *J. Alternative Complementary Med,* 10, 917-918 (2004).
[6] H. Green-Pedersen, B. T. Jensen, N. Pind, *Env.Technol*, 18, 807-815 (1997).
[7] Ch.-K. D. Hsi, D. Langmuir, *Geochim. Cosmochim. Act,* 49, 1931-1941 (1985).
[8] I. A. Katsoyiannis, H. W. Althoff, H. Bartel, M. Jekel, *Water Res*., 40, 3646-3652 (2006).
[9] U. Palmqvist, E. Ahlberg, L. Lovgren, S. Sjoberg, *J. Colloid Interf. Sci*, 218, 388-396 (1999).
[10] M. Wazne, G. P. Korfiatis, X. Meng, *Env. Sc. Technol*., 37, 3619-3624 (2003).