ISSN 1070-3632, Russian Journal of General Chemistry, 2013, Vol. 83, No. 6, pp. 1299–1304. © Pleiades Publishing, Ltd., 2013.
Original Russian Text © D. M. Zhilin, 2011, published in Rossiiskii Khimicheskii Zhurnal, 2011, Vol. 55, Nos. 5–6, pp. 95–99.

Measurement of the Amount of Information Contained in a Chemical Equation (Preliminary Estimation)

D. M. Zhilin

Moscow Institute of Open Education, Aviatsionnyi per. 6, Moscow, 125167 Russia
e-mail: zhila2000@mail.ru

Received April 24, 2011

Abstract—The entropy of chemical language was estimated basing on Shannon's theory of information. The alphabet of the language was defined. It contains 110 symbols of chemical elements, indices, coefficients, brackets, condition signs, and some other (a total of 166 symbols). The probabilities of the symbols were calculated using a higher school chemistry textbook. The upper bound of entropy of chemical language was estimated at 4.55 bits per symbol. The table of frequencies and self-information for the language was given, and the way to calculate the amount of information was shown.

DOI: 10.1134/S1070363213060509

Secondary and higher school teachers tend to judge chemical equations intuitively as "complex" or "simple." For example, equation

$$NaOH + HCl = NaCl + H_2O$$

is intuitively "simpler" than equation

$$I_2 + 3KIO_3 + 18HCl + 2KCl = 5K[ICl_4] + 9H_2O.$$

In this connection, a question arises, how can the complexity of a chemical equation be measured quantitatively? This question is not idle, because a quantitative measure of complexity of chemical equations can (and should) be applied to both organizing and managing the teaching process and assessing the learning outcomes.

First thing that needs to be mentioned is that "going from simple to complex" implies ranking objects (in particular, chemical equations in chemistry teaching case) based on their complexity. This principle was set forth in modern cognitive theories, e.g., chunking theory [1, 2] and cognitive load theory [3, 4].

Chunk (a term introduced into theory [1, 2] and often used in Russian-language literature) is an object stored in long-term memory, which is manipulated by the working memory as a unified whole. Specifically chunking allows resolving the fundamental contradiction between the very limited capacity of working memory (seven plus or minus two objects [5]) and huge complexity of objects cognizable by human mind. Working memory can actually manipulate a limited number of chunks, but chunks themselves can be arbitrarily complex. Hence, the learning process can be interpreted as chunking.

It is reasonable that the complexity of chunks be measured a priori. Chemical equations are suitable as chunk components and, consequently, their complexity should be measured a priori.

Cognitive loading is the load imposed on the working memory during the process of learning [6]. There exist intrinsic cognitive load (the inherent level of complexity associated with instructional materials) and external cognitive load (generated by the manner in which information is presented to learners, being under the control of instructional designers). For better perception of a material the cognitive loading should be optimized and thus measured a priori. Chemical equations constitute an integral part of instructional materials for teaching and learning chemistry, which contribute to cognitive loading, and their complexity should be estimated in order that they be included in the overall cognitive load of the material.

However, the chunking theory and the cognitive loading theory still remain qualitative. Brünken et al. [7] described eight approaches to measuring the cognitive load, all of which are based on physiological and psychological state of learners and do not allow a priori estimation.

As mentioned above, the measurement of complexity of chemical equations can be essential for assessing the learning outcomes. So far, this task was fulfilled by two theories: the classical test theory and the item response theory. Both theories attempt to eliminate two parameters that could not be defined: learners' abilities and task complexity [8, 9]. The drawback suffered by these theories consists in that they are based on several hidden strong presumptions (e.g., normal distribution of learners' abilities), whose adequacy (and hence the adequacy of their ensuing theories) is questionable. At the same time, one is justified in assuming that the more complex the equations written by learners, the deeper their knowledge of chemistry. Thus, by directly measuring the complexity of equations, which are written or reproduced by learners, it will be possible to measure their knowledge without making strong presumptions. It only remains to find the right tool for quantitative measurement of complexity of chemical equations.

For a chemical equation interpreted as a text written in a chemical language, such tool may be found in Shannon's information theory [10], which was developed long ago and demonstrated its adequacy by sixty years of application. Specifically this theory underlies all modern theories of information transmission, processing, and storage; it is used for measuring the complexity of various texts [11]. However, modern cognitive theories do not employ this tool for some unknown reason, though cognition is nothing else but assimilation and processing of information. This is all the more surprising that Miller [5] determined the number of working memory objects and discussed the processing of information in terms of Shannon's theory and even measured it in bits. Specifically the size of the chunk in bits allows characterizing its complexity, and the amount (in bits) of specific information in instructional materials can determine their cognitive load. Hence, the amount of information contained in a chemical equation interpreted as a text also determines its complexity.

The amount of information contained in a text can be estimated basing on Shannon's theory [12].

Shannon's theory treats information as a measure of reduced uncertainty: One bit of information means reduction of uncertainty by 50 percent. Presumably, there is a text consisting of two symbols (e.g., 0 and 1). Before the next symbol is read, it can be presumed that it is either zero or one. On reading the next symbol, a choice is made between two alternatives, i.e., the uncertainty is reduced by 50%. Accordingly, each of the symbols of a two-symbol language carries one bit of information. In the case of a four-symbol alphabet (like, e.g., in genetic code), selection of one of the symbols reduces the uncertainty four times (two times by 50%). Hence, one symbol of a four-symbol alphabet carries 2 bits of information. The general formula for the entropy in bits per symbol I for a language whose alphabet consists of N symbols is expressed by the logarithm of the number of symbols to the base 2:

$$I = \log_2 N. \quad (1)$$

However, this formula is valid only if all symbols in the language are equally probable. Indeed, in guessing Russian alphabet letters (like, e.g., in the "Field of Dreams" game), Russian letter "о" will be guessed with a much higher probability than Russian letter "ъ" (hard sign). Hence, guessing letter "о" means reducing the uncertainty to a lesser extent compared to hard sign. Therefore, for alphabets consisting of nonequally probable symbols the amount of information carried by ith symbol (self-information) will be determined by the probability of the symbol in the language p_i:

$$I_i = -\log_2 p_i. \quad (2)$$

This formula implies that, the higher the frequency of this symbol in the alphabet, the smaller the amount of information it carries. As to the weighted average of the self-information of each of the various symbols (entropy), it can be calculated by the formula

$$I = -\sum_i p_i \log_2 p_i. \quad (3)$$

In order to apply this approach to chemical equations, it is necessary to (a) establish the alphabet of chemical equations and (b) calculate the frequency for each symbol i in it.

Analysis of chemical equations available in literature (if they are written with the use of only molecular formulas, and there are no structural formulas) reveals the following symbols of chemical alphabet:

Symbols of the elements (Na, Cl, etc.): 112.

Indices: theoretically unlimited in number; practically rarely exceeds 12.

Brackets (round and square): 4.

Coefficients: theoretically unlimited in number; practically rarely exceed 20. They always follow the

Table 1. Frequency table of the chemical language symbols, compiled based on [13]

Symbol	Frequency	Probability p_i	$-p_i \times \log_2 p_i$	Self-information per symbol,[a] bit
O	960	0.1511	0.412	2.73
$_2$ (index)	856	0.1347	0.390	2.89
H	675	0.1062	0.344	3.23
+2	362	0.0570	0.236	4.13
+	357	0.0562	0.233	4.15
=	355	0.0559	0.233	4.16
$_3$ (index)	285	0.0449	0.201	4.48
S	227	0.0357	0.172	4.81
$_4$ (index)	217	0.0342	0.166	4.87
Cl	211	0.0332	0.163	4.91
N	179	0.0282	0.145	5.15
Na	147	0.0231	0.126	5.43
C	135	0.0212	0.118	5.56
K	112	0.0176	0.103	5.83
+3	106	0.0167	0.099	5.91
–	90	0.0142	0.087	6.14
(	83	0.0131	0.082	6.26
)	83	0.0131	0.082	6.26
+4	75	0.0118	0.076	6.40
Ca	70	0.0110	0.072	6.50
t° (condition)	67	0.0105	0.069	6.57
Si	50	0.0079	0.055	6.99
F	38	0.0060	0.044	7.39
Zn	36	0.0057	0.042	7.46
Fe	35	0.0055	0.041	7.50
–	35	0.0055	0.041	7.50
+6	31	0.0049	0..037	7.68
I	31	0.0049	0.037	7.68
Sn	31	0.0049	0.037	7.68
Mn	29	0.0046	0.035	7.78
P	28	0.0044	0.034	7.83
Cu	26	0.0041	0.032	7.93
+5	24	0.0038	0.030	8.05
As	23	0.0036	0.029	8.11
Mg	22	0.0035	0.028	8.17
[	19	0.0030	0.025	8.39
]	19	0.0030	0.025	8.39

Table 1. (Contd.)

Symbol	Frequency	Probability p_i	$-p_i \times \log_2 p_i$	Self-information per symbol,[a] bit
D	18	0.0028	0.024	8.46
Ag	18	0.0028	0.024	8.46
Cr	18	0.0028	0.024	8.46
Pb	18	0.0028	0.024	8.46
Al	16	0.0025	0.022	8.63
$_6$ (index)	14	0.0022	0.019	8.83
Br	14	0.0022	0.019	8.83
+8	13	0.0020	0.018	8.93
B	12	0.0019	0.017	9.05
Ba	12	0.0019	0.017	9.05
$_7$ (index)	11	0.0017	0.016	9.17
$_5$ (index)	8	0.0013	0.012	9.63
Bi	8	0.0013	0.012	9.63
×	4	0.0006	0.007	10.63
+10	4	0.0006	0.007	10.63
Be	4	0.0006	0.007	10.63
Ge	4	0.0006	0.007	10.63
Ti	4	0.0006	0.007	10.63
Xe	4	0.0006	0.007	10.63
+7	3	0.0005	0.005	11.05
+12	2	0.0003	0.004	11.63
$_{17}$ (index)	2	0.0003	0.004	11.63
$_{35}$ (index)	2	0.0003	0.004	11.63
Au	2	0.0003	0.004	11.63
Pt	2	0.0003	0.004	11.63
$_8$ (index)	1	0.0002	0.002	12.63
$_{12}$ (index)	1	0.0002	0.002	12.63
7	1	0.0002	0.002	12.63
p (condition)	1	0.0002	0.002	12.63
+11	1	0.0002	0.002	12.63
+16	1	0.0002	0.002	12.63
+18	1	0.0002	0.002	12.63

[a] Calculated by formula (2).

"=" or "+" sign, with "+" sign characterized by a higher frequency. Therefore, the "+" sign can be conjoined with the succeeding coefficient.

Sign "+", if there is no succeeding coefficient, i.e., if the coefficient is presumed to be equal to unity.

Charge signs: theoretically unlimited number; practically rarely go beyond ± 4.

Point, like, e.g., in $CuSO_4 \cdot 5H_2O$.

Signs ↓ (precipitate) and ↑ (gas): 2.

Symbols designating the heat effect (+*Q* and –*Q*): 2. The specific value of the heat effect carries its intrinsic information and thus can be eliminated from the chemical alphabet.

Equality sign ("="). If this symbol is followed by a coefficient, it is considered as one symbol together with the corresponding coefficient succeeding the "+" sign, conjoined with this sign.

Symbols designating the reaction conditions (usually placed above the equality sign): theoretically unlimited in number; practically often equal six: reversibility ($\rightleftarrows$), heating ($t°$), pressure (p), light ($h\nu$), electrolysis ($\downarrow$), and catalyst (cat). These symbols carry information, additional to that carried by symbol "=". If the catalyst formula is provided, the corresponding symbols and indices also contribute to the overall amount of the information contained in the equation. If particular values are given for temperature and pressure, they carry their intrinsic information and can be eliminated from the chemical alphabet.

Overall, there are 166 symbols. If they were equally probable, there would be $I = \log_2 166 = 7.4$ bits per symbol on the average. However, the symbols have different frequencies. To determine the frequency of each symbol (or, to draw a frequency table) we used the classical textbook by Glinka [13]. It contains 355 equations in which all symbols were counted. Table 1 presents the results.

It is seen that, out of the 166 chemical language symbols, only 69 were used in textbook [13], which would give the entropy of 6.1 bits per symbol, if the symbols were equally probable. Calculations with due regard to nonequal probability of the symbols [by formula (3)] give a value reduced to 4.5 bits per symbol.

The use of Table 1 for calculating the amount of information contained in a specific equation can be illustrated by the following example:

$$2SO_2 + O_2 \overset{t°,\ p,\ \mathrm{Pt}}{\rightleftarrows} 2SO_3. \qquad (4)$$

A table of the symbols used is to be compiled in which each symbol will be provided with the self-information data (the last column of Table 1), and the resulting values will be added together (see Table 2).

As a result, Eq. (4) as it is written contains 83.9 bits of information. Interestingly, if the reaction conditions (reversibility, heating, pressure, and platinum catalyst) will be eliminated, the amount of information will be reduced to 44.6 bits, e.g., nearly by half.

The procedure proposed here allows only preliminary estimation of the amount of information contained in a chemical equation. To obtain a more accurate estimate, it seems reasonable to process the reaction equations given in other textbooks. Apparently, it makes sense to turn to higher-level textbooks to include rare symbols. To obtain a close to universal table, it is necessary to determine how the order of symbols changes with textbook. Furthermore, there is a need to somehow take into account symbols of conditions which are not specified in equations but referred to in the text. For example, the chemical

Table 2. Results of calculation of the amount of information contained in chemical equation (4)

Symbol	Number of symbols	Self-information per symbol,[a] bit	Self-information of all symbols, bit
+2	2	4.13	8.26
S	2	4.81	9.62
O	3	2.73	8.19
$_2$	2	2.89	5.78
+	1	4.15	4.15
=	1	4.16	4.16
D	1	8.46	8.46
$t°$	1	6.57	6.57
p	1	12.63	12.63
Pt	1	11.63	11.63
$_3$	1	4.48	4.48
Total			83.93

[a] The last column of Table 1.

equations in textbook [13] are not supplied with symbols of electrolysis and catalyst (these conditions are only mentioned in the text), and the corresponding symbols are lacking in the frequency table. Finally, there is a need to find a way to deal with the symbols provided after the equality sign so that the way in which addends are grouped does not change the amount of information contained in the equation.

Nevertheless, the preliminary estimation procedure described here allows quantitative assessment of the complexity of different chemical equations and their intercomparison. This opens prospects for development of materials for teaching chemistry and assessing learning outcomes with the use of the information theory apparatus.

REFERENCES

1. Chase, W.G. and Simon, H.A., *Perception in Chess, Cognit. Psychol*., 1973, vol. 1, no. 4, p. 55–81.
2. Sweller, J., *Evolution of Human Cognitive Architecture*, in *The Psychology of Learning and Motivation*, Ross, B., Ed., San Diego: Academic, 2003, vol. 43, pp. 215–266.
3. Chandler, P., Cooper, G., Pollock, E., and Tindall-Ford, S., *Applying Cognitive Psychology Principles to Education and Training*, 1998; *http://www.aare.edu.au/98pap/cha98030.htm.*
4. Paas, F., Renkl, A., and Sweller, J., *Educ. Psychol*., 2003, vol. 38, no. 1, pp. 1–4.
5. Miller, J.A., *Psychol. Rev*., 1956, vol. 101, no. 2, pp. 343–352.
6. Paas, F. and van Merriënboer, J.J.G., *Educ. Psychol. Rev*., 1994, vol. 6, no. 1, pp. 51–71.
7. Brünken, R., Plass, J.L., and Leutner, D., *Educ. Psychol*., 2003, vol. 38, no. 1, pp. 53–61.
8. Hambleton, R.K. and Jones, R.W., *Educ. Meas.: Issues Practice*, 1993, vol. 12, no. 3, pp. 535–556.
9. Neiman, Yu.M. and Khlebnikov, V.A., *Vvedenie v teoriyu modelirovaniya i parametrizatsii pedagogicheskikh testov* (Introduction to Modeling and Parameterization of Pedagogic Tests), Moscow: Prometei, 2000.
10. Shannon, C.E., *The Bell Syst. Tech. J.,* 1951, vol. 30, pp. 50–64.
11. Piotrovskii, R.G*., Informatsionnye izmereniya yazyka* (Information Measurement of Language), Leningrad: Nauka, 1968.
12. Yaglom**,** A.M. and Yaglom**,** I.M**.,** *Probability and Information*, Norwell: Reidel**,** 1983**.**
13. Glinka, N.L., *Obshchaya khimiya* (General Chemistry), Leningrad: Khimiya, 1984.