How Much Information Do Chemical Equations Contain?

Denis M. Zhilin
Moscow Institute for Open Education; School 192, Moscow, Russia

Abstract

The entropy of "tertiary students" chemical language is estimated. The alphabet of the language contains 110 symbols of chemical elements, indexes, coefficients, brackets, condition signs and some other (totally 167 symbols). The probabilities of the symbols were calculated using a textbook on chemistry for tertiary institution. The upper bound of entropy of chemical language is estimated to be 4.55 bits/symbol that could be decreased to 4.05. The table of frequencies and self-information for the language is given.

Key words: cognition, information theory, entropy, chemical language, chemical equations.

Synopsis

Innovative learning requires adequate organisation of cognition process and information within it. This problem has been studying for last several decades. At least two complement cognitive theories have been developed: chunking theory (Chase, Simon, 1973) and theory of cognitive loading (Chandler et. al., 1998). Chunk is referred to long-term memory structures that can be used as a unit of perception and meaning. Operating with rather complex chunks a person handles the limitations of working memory both in capacity and duration. Cognitive loading performs a particular task imposes on the learner's cognitive system (Paas & van Merrienboer, 1994) and should be optimized.

However, both mentioned above theories still remain qualitative. As soon as the chunk is a complex structure, how is it possible to measure its complexity? The cognitive load can "increase" or "decrease", but how to measure, how much? Brunker et al (2003) reviewed eight approaches of measuring cognitive load, but all of them are based on physiological and psychological state of a student and still cannot be calculated basing on the properties of instruction material.

Another important problem is an estimation of abilities or proficiency. Two theories – classical test theory and item response theory – were developed to estimate it basing on test performance (Hambleton, Jones, 1993). Both of them try to eliminate two parameters that could not be defined: ability and complexity of a task. The presumptions of the theories and hence their adequacy are disputable.

Surprisingly none of the theories of cognition (i.e. processing of information) and testing (i.e. retrieval of information) employ classical Shannon theory of information – the foundation of all the innovations for the recent half of a century. It becomes more surprisingly when we turn to the classical article of Miller, 1956 on chunking theory, where he discussed the theory of information and measured transmitted and recalled information in bits. Meanwhile, the amount of information (in bits) in one chunk would reasonable characterise it. The amount (or portion) of excessive information in instructional materials would characterize the redundancy effect that implies cognitive loading. The amount of information to be proceeded to solve a task would objectively characterize a complicity of the task.

It is obvious, that only significant information for a particular domain should be measured. To extract significant information we should make some presumptions about a nature of domain tasks. For chemistry we can presume that its main problem is to predict the result of chemical reaction within certain conditions. The predictions are described in chemical equations, so the significant information for chemistry is contained there. The more information the equation contents, the more complex it is. So, in order to apply information theory to chemistry teaching and learning we should solve a key task: to measure information contained in chemical equations.

For this purpose we could regard chemical equations as a particular language and apply tools, developed by Shannon, 1951. It connects information, containing in one symbol with its entropy H – a measure of unpredictability of the symbol. Entropy (in bits per symbol) of the whole language is defined as $H = \log_2 N$

(N is a number of symbols in the alphabet) when symbols are equally probable and

$$H = -\sum_i p_i \cdot \log_2 p_i$$

(p_i is the probability of the symbol in the language) when symbols are more or less probable. The amount of information, carried by a particular i-th symbol (a self-information) is
$I_i = -\log_2 p_i$.

To apply Shannon's approach to chemical language its alphabet and the frequency of the symbols should be established. The alphabet of chemical language employs the following symbols:

(a) Symbols of the elements (Na, Cl and so on) – 110.
(b) Indexes (theoretically unlimited, but rarely exceeding 12).
(c) Brackets (round and square) – 4.
(d) Coefficients (theoretically unlimited, but rarely exceeds 20). They also follow sign "+" or "=", so sign "+" can be conjoined with the following coefficient.
(e) Charge signs (theoretically unlimited but rarely go beyond ±4).
(f) Point (as in $CuSO_4 \cdot 5H_2O$).

(g) Signs ↓ (precipitate) and ↑ (gas) – 2;
(h) Signs denoting heat effect (+Q and -Q) – 2. The particular value of heat effect carries its own information, so it can be excluded from chemical alphabet.
(i) Equation symbol ("=") and end of equation (there is no symbol, but it is implied)
(j) Symbols denoting conditions. Their number is theoretically unlimited, but the significant are those denoting reversibility, heating, pressure, light, electrolysis and catalyst (totally 6, all of them are added to a simple equation symbol). If the catalyst is specified, its formula contributes to a number of symbols and indexes. If the particular values are mentioned, they carry their own information and can be excluded from the chemical alphabet.

Summarizing we have 167 symbols that gives $I = \log_2 167 = 8$ bits for symbol. However, the frequency of the symbols is quite different. To get a preliminary estimation of it we counted symbols in equations in Russian version of a classical textbook for tertiary institutions by Glinka (1981) with 355 equations. The results are summarized in Table 1.

Table 1.

Frequency of different symbols in chemical alphabet and their entropy

Symbol	Frequency	Probability	$-p_i \cdot \log_2 p_i$	Self-inform.	Symbol	Frequency	Probability	$-p_i \cdot \log_2 p_i$	Self-inform.
O	959	0.1433	0.4016	2.8	3 (coeff.)	105	0.0157	0.094	6.35
2 (index)	857	0.128	0.3797	2.97	↑	90	0.0134	0.0836	6.35
H	676	0.101	0.3341	3.31	(	82	0.0123	0.0778	6.48
I (coeff.)*	362	0.0541	0.2276	4.21	)	82	0.0123	0.0778	6.58
2 (coeff.)	361	0.0539	0.2272	4.21	4 (coeff.)	75	0.0112	0.0726	6.64
=	355	0.053	0.2247	4.24	Ca	70	0.0105	0.0688	7.06
End	355	0.053	0.2247	4.24	heating	67	0.01	0.0665	7.46
3 (index)	285	0.0426	0.1939	4.55	Si	50	0.0075	0.0528	7.54
S	228	0.0341	0.1661	4.88	F	38	0.0057	0.0424	7.58
4 (index)	217	0.0324	0.1604	4.95	Zn	36	0.0054	0.0405	7.58
Cl	211	0.0315	0.1572	4.99	Fe	35	0.0052	0.0396	7.75
N	178	0.0266	0.1392	5.23	B	12	0.0018	0.0164	7.75
Na	147	0.022	0.121	5.51	7 (index)	11	0.0016	0.0152	7.8
↓	35	0.0052	0.0396	5.64	Sn	31	0.0046	0.0359	7.85
1	31	0.0046	0.0359	5.9	6 (coeff.)	30	0.0045	0.035	7.9
C	134	0.02	0.113	5.99	Mn	29	0.0043	0.034	8.01
K	112	0.0167	0.0987	6.22	P	28	0.0042	0.0331	8.12

Table 1 (continue)

Symbol	Frequency	Probability	$-p_i \cdot \log_2 p_i$	Self-inform.	Symbol	Frequency	Probability	$-p_i \cdot \log_2 p_i$	Self-inform.
Cu	26	0.0039	0.0311	8.18	10 (coeff.)	4	0.0006	0.0064	10.7
5 (coeff.)	24	0.0036	0.0291	8.25	Be	4	0.0006	0.0064	10.7
As	23	0.0034	0.0281	8.46	Ge	4	0.0006	0.0064	10.7
Mg	22	0.0033	0.0271	8.46	Ti	4	0.0006	0.0064	10.7
]	19	0.0028	0.024	8.54	Xe	4	0.0006	0.0064	10.7
[	19	0.0028	0.024	8.54	7 (index)	3	0.0004	0.005	11.1
⇆	19	0.0028	0.024	8.54	12 (index)	2	0.0003	0.0035	11.7
Cr	18	0.0027	0.023	8.54	17 (index)	2	0.0003	0.0035	11.7
Pb	18	0.0027	0.023	8.54	35 (index)	2	0.0003	0.0035	11.7
Ag	18	0.0027	0.023	8.8	8 (index)	1	0.0001	0.0019	12.7
Al	16	0.0024	0.0208	8.9	12 (index)	1	0.0001	0.0019	12.7
6 (index)	15	0.0022	0.0197	9.01	electrol.	1	0.0001	0.0019	12.7
Br	14	0.0021	0.0186	9.12	pressure	1	0.0001	0.0019	12.7
8 (coeff.)	13	0.0019	0.0175	9.25	11 (coeff.)	1	0.0001	0.0019	12.7
5 (index)	8	0.0012	0.0116	9.71	16 (coeff.)	1	0.0001	0.0019	12.7
Bi	8	0.0012	0.0116	9.71	18 (coeff.)	1	0.0001	0.0019	12.7
.	4	0.0006	0.0064	10.7					

* coefficient "1" is omitted in real equations.

Only 67 types of symbols were used in the textbook. The total number of symbols was 6693 and the entropy turned to be 4.55 bits/symbol. However, this estimation refers to "tertiary students" chemical language. If we analyze, for example "Inorganic chemistry" by Remy we shall find more symbols and the frequency will be different. Really the results should be considered as preliminary – further we shall investigate a larger set of textbooks.

The value of 4.55 bits/symbol is the upper bound. Constructing "random equations" (randomly placing symbols one after another according to their probability) we came across senseless combinations (such as index after coefficient or two same symbols in one "substance"). Only 181 random symbols of 256 (71%) can be met in the real equation after previous chain of symbols. That reduces the estimated upper bound of the entropy to 4.05 bits/symbol. Analysing the sequence of the symbols would decrease the upper bound more.

The results on self-information allow calculating the information contained in any chemical equation thus comparing the complexity of equations. It is sufficiently to sum up the values of self-information of all symbols in the equation.

References

Brunker R., Plass J.L., Leutner D. (2003). Direct Measurement of Cognitive Load in Multimedia Learning. *Educational Psychologist.* 38 (1) 53 - 61.

Chandler P., Cooper G., Pollock E., Tindall-Ford S. (1998). Applying Cognitive Psychology Principles to Education and Training. http://www.aare.edu.au/98pap/cha98030.htm.

Chase, W. G., Simon H.A. (1973). Perception in chess. *Cognitive Psychology.* 4 (1) 55-81

Glinka N. (1981) General Chemisty. Moscow: Mir Publishers.

Hambleton, R. K., Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: issues and practice.* 12 (3) 535-556.

Miller J.A. (1956) The Magical Number Seven, Plus or Minus Two. Some Limits on Our Capacity for Processing Information. *Psychological Review.* 101 (2) 343-352.

Paas, F., van Merriënboer, J. J. G. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review.* 6 (1) 51–71.

Shannon C.E. (1951). Prediction and Entropy of Printed English. *The Bell System Technical Journal.* 30 50-64