Измерение количества информации в химическом уравнении (предварительная оценка)

Д. М. Жилин

ДЕНИС МИХАЙЛОВИЧ ЖИЛИН — кандидат химических наук, доцент Московского института открытого образования и преподаватель средней общеобразовательной школы № 192 (Москва). E-mail: zhila2000@mail.ru

Учителя школ и преподаватели вузов часто судят о химических уравнениях на интуитивном уровне в терминах «сложность» и «простота». Так, на интуитивном уровне очевидно, что уравнение

$$NaOH + HCl = NaCl + H_2O$$

«проще», чем уравнение

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

Возникает вопрос: как можно измерить сложность химических уравнений количественно. Вопрос этот далеко не праздный, ибо количественную меру сложности химических уравнений можно (и нужно) использовать как при организации обучения, так и в контроле знаний учащихся.

Начнём с того, что принцип «от простого к сложному» предполагает некоторое ранжирование объектов (в том числе химических уравнений при преподавании химии) по сложности. Этот принцип раскрывается в таких современных когнитивных теориях, как теория пластов познания (chunking theory [1, 2]) и теория познавательной нагрузки (theory of cognitive loading, [3, 4]).

Пласт познания (в русскоязычной литературе часто используют это название для понятия chunk, введенного в теорию [1, 2]) представляет собой объект в долговременной памяти, которым рабочая память оперирует как целым. Именно организация объектов в пласты позволяет разрешить фундаментальное противоречие между крайне ограниченной ёмкостью рабочей памяти (семь плюс—минус два объекта, [5]) и огромной сложностью объектов, которые способен понять человек. Рабочая память действительно может оперировать с ограниченным числом пластов, но сами пласты могут быть сколь угодно сложными. Соответственно, процесс обучения представляется как процесс формирования пластов познания.

Сложность формируемых пластов познания целесообразно измерять априорно. Поскольку химические уравнения вполне могут быть компонентами пластов познания, априорно следует измерять сложность уравнений.

Познавательная нагрузка (cognitive loading) есть нагрузка на рабочую память в процессе познания [6]. Она делится на нагрузку, присущую материалу (intrinsic) и внешнюю нагрузку, зависящую от того, как организован материал (external). Для наилучшего восприятия материала познавательная нагрузка должна быть оптимизирована, а значит, измерена априорно. Поскольку химические уравнения — неотъемлемая часть обучающих материалов по химии, вносящих свой вклад в познавательную нагрузку, следует оценивать их сложность, чтобы включить их в общую познавательную нагрузку материала.

Однако до сих пор теория пластов познания, равно как и теория познавательной нагрузки, остаются качественными. Описано [7] восемь подходов к измерению познавательной нагрузки, и все они базируются на физиологическом и психологическом состоянии студентов в процессе обучения и никак не позволяют провести априорную оценку.

Как указано выше, измерение сложности химических уравнений может пригодиться при оценке достижений учащихся. До сих пор в этих целях используются две теории: классическая теория тестирования (Classical Test Theory) и теория ответов на отдельные вопросы (Item Response Theory). Обе они направлены на то, чтобы сократить два неопределяемых параметра — способность учащегося и сложность задания [8, 9]. Недостаток этих теорий состоит в том, что в их основе скрыто несколько достаточно сильных предположений (например, о нормальности распределения учащихся по способностям), адекватность которых (а значит,

и адекватность следующих из них теорий) сомнительна. Между тем можно принять, что чем более сложные уравнения способен написать учащийся, тем глубже его познания в химии. Тем самым прямое измерение сложности уравнений, которые пишет или воспроизводит учащийся, позволит измерить его познания, не прибегая к сильным предположениям. Осталось найти подходящий инструмент для количественного измерения сложности химических уравнений.

Если рассматривать химическое уравнение как текст, составленный с использованием некого химического языка, то такой инструмент уже давно разработан. Это теория информации К. Шеннона [10], которая за прошедшие с момента её разработки шестьдесят лет показала свою адекватность. Именно на основе этой теории построены все современные системы передачи, обработки и хранения информации. Именно с её помощью измеряется сложность тех или иных текстов [11]. Однако почему-то современные теории познания этот инструмент не используют, хотя познание есть не что иное, как усвоение и переработка информации. Это тем более удивительно, что Миллер в своей статье [5] (в которой он и определил число объектов рабочей памяти) обсуждал процесс переработки информации в терминах теории Шеннона и даже измерял её в битах. Именно размер пласта познания в битах позволяет характеризовать его сложность, а количество (в битах) определённой информации в инструктивных материалах может определять их познавательную нагрузку. Значит, и количество информации в химическом уравнении, как в разновидности текста, определяет его сложность.

Как же измерить количество информации, содержащееся в тексте? Ответ на этот вопрос дан в книге [12], и он базируется на теории Шеннона.

По Шеннону, информация есть мера снятой неопределённости. Один бит информации есть снижение неопределённости в два раза. Предположим, имеется текст, состоящий из двух символов (например, 0 и 1). Пока мы не прочитали следующий символ, можно принять, что он будет либо нулем, либо единицей. Как только мы этот символ прочитали, мы сделали один выбор из двух возможных, т. е. уменьшили неопределённость в два раза. Соответственно, один символ двусимвольного языка несёт один бит информации. Если символов в языке четыре (как, например, в генетическом коде), то выбор одного из них снижает неопределённость в четыре раза, т. е. в дважды по два раза. Значит, один символ четырехсимвольного языка несёт 2 бита информации. Общая формула для количества информации I в языке, алфавит которого состоит из N символов, в расчете на один символ выражается логарифмом от числа символов по основанию 2:

$$I = \log_2 N \qquad (1)$$

Однако эта формула справедлива только в том случае, если все символы в языке равновероятны. В самом деле, если мы будем отгадывать буквы русского текста (как в «Поле чудес»), то гораздо вероятнее, что мы угадаем букву «о», чем букву «ъ». Значит, угадав букву «о», мы снимем неопределённость в меньшей степени, чем угадав букву «ъ». Поэтому для языков с неравновероятными символами количество информации, которую несёт i-й символ (его статистический вес), определяется вероятностью его встречаемости в языке p_i :

$$I_i = -\log_2 p_i \qquad (2)$$

Из этой формулы следует, что чем чаще встречается символ в алфавите, тем меньше информации он несёт. Что касается средней информации в расчёте на один символ, то она вычисляется по формуле

$$I = -\sum_i p_i \cdot \log_2 p_i \qquad (3)$$

Чтобы применить этот подход к химическим уравнениям нужно: (а) сформировать алфавит языка химических уравнений и (б) посчитать вероятность каждого символа в нём.

Анализ имеющихся в литературе химических уравнений (при условии, что в них используются только формулы состава веществ и не записаны структурные формулы) позволяет выделить следующие символы химического алфавита.

- Символы элементов (Na, Cl и т.п.) — 112.
- Индексы, теоретически их число неограничено, однако практически редко превышает 12.
- Скобки (круглые и квадратные) — 4.
- Коэффициенты, теоретически их число неограничено, но практически они редко превышают 20. Они всегда следуют за знаками «=» или «+», причем знак «+» встречается чаще. Поэтому знак «+» вместе со следующим за ним коэффициентом можно считать за один символ.
- Знак «+», если за ним не следует коэффициент, т.е. подразумевается коэффициент, равный 1.

- Знаки зарядов, теоретически их число неограничено, однако практически они редко выходят за пределы ±4.
- Точка, как, например, в $CuSO_4 \cdot 5H_2O$.
- Знаки ↓ (осадок) и ↑ (газ) – 2.
- Знаки, обозначающие тепловой эффект (+*Q* и –*Q*) — 2. Точное значение теплового эффекта несёт собственную информацию, поэтому может быть исключено из химического языка.
- Символ равенства («=»). Если после этого символа стоит коэффициент, то он считается тем же символом, что и соответствующий коэффициент, следующий за знаком «+» вместе с этим знаком.
- Символы, обозначающие условия протекания реакции (обычно размещаются над знаком равенства). Теоретически их количество неограничено, однако практически часто встречаются шесть символов: обратимость «⇌», нагревание «*t*°», давление «*p*», свет «*h*ν», электролиз «ϟ» и катализатор «кат». Эти символы несут дополнительную информацию к символу «=». Если указана формула катализатора, то записанные в ней символы и индексы вносят свой вклад в общую информацию, содержащуюся в уравнении. Если указаны точные значения температуры и давления, то они несут свою информацию и из химического языка могут быть исключены.

Итого получается 166 символов. Если бы они были равновероятны, то в среднем на один символ приходилось бы $I = \log_2 166 = 7{,}4$ бит. Однако символы неравновероятны. Чтобы определить вероятность каждого символа (или, как ещё говорят, составить частотную таблицу), мы воспользовались классическим учебником Глинки [13]. Он содержит 355 уравнений, в которых были посчитаны все символы. Результаты представлены в табл. 1.

Из таблицы видно, что из 166 символов химического языка в учебнике Глинки было ис-

Таблица 1

Частотная таблица символов химического языка, составленная по [13]

Символ	Количество встреч	Вероятность p_i	$-p_i \log_2 p_i$	Информация в символе*, бит
O	960	0,1511	0,412	2,73
$_2$ (индекс)	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
$_3$ (индекс)	285	0,0449	0,201	4,48
S	227	0,0357	0,172	4,81
$_4$ (индекс)	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° (условие)	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

Таблица 1 (Продолжение)

Символ	Количество встреч	Вероятность p_i	$-p_i \cdot \log_2 p_i$	Информация в символе*, бит
+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
⇆	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
$_6$ (индекс)	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
$_7$ (индекс)	11	0,0017	0,016	9,17
$_5$ (индекс)	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
$_{17}$ (индекс)	2	0,0003	0,004	11,63
$_{35}$ (индекс)	2	0,0003	0,004	11,63
Au	2	0,0003	0,004	11,63
Pt	2	0,0003	0,004	11,63
$_8$ (индекс)	1	0,0002	0,002	12,63
$_{12}$ (индекс)	1	0,0002	0,002	12,63
↯	1	0,0002	0,002	12,63
p (условие)	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

* Рассчитана по формуле (2).

пользовано всего 69, что дало бы 6,1 бит/символ, если бы символы были равновероятны. Однако учет неравновероятности [расчёт по формуле (3)] снижает эту величину до 4,5 бит/символ.

Как пользоваться этой таблицей, если требуется рассчитать количество информации, которое содержится в том или ином уравнении? Приведём пример:

$$2SO_2 + O_2 \overset{t^\circ, p, \mathrm{Pt}}{\rightleftharpoons} 2SO_3 \qquad (4)$$

Составляют таблицу использованных символов, приписывают каждому, сколько информации он содержит (из последней колонки табл. 1) и суммируют их (табл. 2).

Таблица 2

Результаты расчета количества информации в химическом уравнении (4)

Символ	Количество символов	Информация в одном символе*, бит	Информация во всех символах, бит
+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
⇆	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
Итого			83,93

* Последний столбец табл. 1.

Получается, что в уравнении (4) в том виде, как оно записано, содержится 83,9 бит информации. Интересно, что если убрать из него указания на условия протекания реакции (обратимость, нагревание, давление и платиновый катализатор), то количество информации уменьшится до 44,6 бит, т. е. почти в два раза.

Предложенная оценка количества информации в химическом уравнении носит только предварительный характер. Для более точной оценки представляется целесообразным обработать уравнения реакции, приведенные в других учебниках. Видимо, имеет смысл обратиться к учебникам более высокого уровня, чтобы охватить редко встречающиеся символы. Чтобы получить таблицу, близкую к универсальной, нужно проследить, как меняется порядок символов при использовании разных учебников. Кроме того, придётся как-то учитывать символы условий, не отмеченных в уравнениях, но упомянутых в тексте. Так, в учебнике [13] в химических уравнениях не указываются обозначения электролиза и катализатора (эти условия упомянуты лишь в описании), поэтому соответствующих символов не оказалось в частотной таблице. И, наконец, нужно решить, как учитывать символы после знака равенства, чтобы количество информации в уравнении не изменялось от перестановки в нём слагаемых.

Тем не менее представленная предварительная оценка уже позволяет количественно оценивать сложность химических уравнений и сравнивать их между собой, что открывает перспективы для разработки учебных и контрольных материалов по химии с использованием аппарата теории информации.

ЛИТЕРАТУРА

1. *Chase, W. G., Simon H.A.* Perception in chess. Cognitive Psychology, 1973, v. 1. № 4, p. 55—81.
2. *Sweller J.* Evolution of Human cognitive architecture. In The Psychology of learning and motivation. San Diego: Academic Press.2003, v. 43, p. 215—266,
3. *Chandler P., Cooper G., Pollock E., 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., Sweller J.* Cognitive Load Theory and Instructional Design: Recent Developments. Educational Psychologist, 2003, v. 38. № 1, p. 1—4.
5. *Miller J.A.* The Magical Number Seven, Plus or Minus Two. Some Limits on Our Capacity for Processing Information. Psychological Review, 1956, v. 101, № 2, p. 343—352.
6. *Paas, F., van Merrienboer, J. J. G.* Instructional control of cognitive load in the training of complex cognitive tasks. Educational Psychology Review, 1994, v. 6, № 1, p. 51—71.
7. *Brunker R., Plass J.L., Leutner D.* Direct Measurement of Cognitive Load in Multimedia Learning. Educational Psychologist. 2003, v. 38, № 1, p. 53—61.
8. *Hambleton, R. K., Jones, R. W.* Comparison of classical test theory and item response theory and their applications to test development. Educational Measurement: issues and practice. 1993, v. 12, № 3, p. 535—556.
9. *Нейман Ю.М., Хлебников В.А.* Введение в теорию моделирования и параметризации педагогических тестов. М.: Прометей, 2000.
10. *Shannon C.E.* Prediction and Entropy of Printed English. The Bell System Technical Journal, 1951, v. 30, p. 50—64.
11. *Пиотровский Р.Г.* Информационные измерения языка. Л.:Наука, 1968.
12. *Яглом А.М., Яглом И.М.* Вероятность и информация. М.: УРСС, 2000.
13. *Глинка Н.Л.* Общая химия. Л.: Химия, 1984.