Национальная академия наук Беларуси

СБОРНИК ТРУДОВ МОЛОДЫХ УЧЕНЫХ НАЦИОНАЛЬНОЙ АКАДЕМИИ НАУК БЕЛАРУСИ

Том V

Минск
"Логвинов"
2004

УДК 655.512

ОПТИМИЗАЦИЯ СЛОЖНОСТИ ТЕКСТА

Ю. Ф. Шпаковский

Белорусский государственный технологический университет, г. Минск, yury_s@tut.by

The relations between words, it's influence on the understanding of text are analyzed. Data about average sentence-length in the chemical publications are cited.

В современном обществе информация играет важную роль в жизни каждого человека. Без нее трудно представить принятие правильных решений, от которых зависит успех любого дела.

Часто возникают ситуации, когда читатель не получает информацию или получает ее не в полном объеме. Попытка понять, почему так происходит, представляет несомненный интерес для исследователей.

При чтении взаимодействуют два объекта: читатель и текст. Каждый из объектов обладает рядом характеристик. Для текста можно выделить такие параметры, как структура, длина предложений, словарный состав и др. Для читателя следует учитывать уровень образования, объем оперативной памяти, уровень предварительных знаний и т. д. Специалисты считают, что при оптимальном сочетании основных характеристик, происходит наиболее полное и адекватное понимание текста (получение информации).

В нашем исследовании мы попытаемся проанализировать влияние структуры предложений на понимание материала с учетом оперативной памяти студентов.

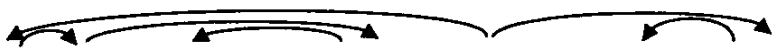
Исследователи трактуют понимание текста как а) понимание слов, т. е. образование связей между словами и отображениями действительности, б) образование связей между словами, в) образование связей между фразами, г) образование связей между частями текста.

Таким образом, понимание есть осознание связей. Приведем пример:

Метод Гиллеспи полезен для объяснения и предсказания строения различных молекул.

Для полного понимания предложения необходимо осознание всех связей.

Анализ формул читабельности (предсказывают трудность того или иного текста для того или иного читателя) показал, что одним из наиболее часто встречающихся показателей трудности текста является длина предложения. Это, по-видимому, связано с тем, что более длинное предложение предполагает осознание большего количества связей. Однако эксперименты показывают, что предложения одинаковой длины могут представлять различную трудность для читателя. Это связано с различными факторами. Мы попытаемся проанализировать характер связей. Приведем пример:



Молекула в невозбужденном состоянии имеет минимальную энергию.

Очевидно, что для понимания межсловесной связи эти слова должны запоминаться одновременно. Если между ними находятся еще слова, то в памяти необходимо удерживать все другие слова. Мы видим, что в предыдущем примере макси-

мальное количество слов, которое необходимо запомнить для образования самой длинной связи равно 3. В последнем примере для образования связи между словами *имеет* и *молекула* необходимо запомнить 5 слов. Таким образом, возникает вопрос: Какое может быть предельное расстояние (или "глубина" связи) между словами, чтобы связь образовалась с первого раза чтения?. Очевидно, что это связано с оперативной памятью читателя.

Исследователями памяти установлено, что объем оперативной памяти человека составляет 7±2 слов [1]. Является ли эта цифра абсолютной для всех? Мы решили исследовать объем оперативной памяти студентов высших курсов химической специальности БГТУ. Память школьников (10 класса) исследовал Я. А. Микк [2].

Итак, для понимания текста следует понять межсловесные связи, которые понимаются тем труднее, чем больше слов расположено между связанными словами. Задача автора и редактора – устранить связи, которые затрудняют понимание, т. е. осознаются не с первого раза чтения. Но для этого необходимо знать предельную трудность связи. Вслед за Я. А. Микком [2] назовем трудностью связи слова количество слов, которые следует запомнить, чтобы образовалась эта связь. Очевидно, что одно слово может быть связано со многими словами, прочитанными раньше. Если запомнить самое левое слово, которое связано с данным словом, то, видимо, запомнятся и все остальные. Таким образом, имеет смысл рассматривать трудность только самой длинной назадидущей связи слова. Назовем ее *трудностью связей слова.*

Теперь определим предельную допустимую трудность связей слова. Но слова в предложении связаны грамматически и семантически. Таким образом, необходимо выяснить, сколько студенты могут запомнить несвязанных и сколько связанных. Было поставлено 2 эксперимента.

Эксперименты проводились на 3 и 5 курсах БГТУ (студенты химической специальности). Ниже приводятся результаты экспериментов только для 3 курса, в котором участвовало 20 человек.

Эксперимент 1. Выясним, сколько несвязанных слов может запомнить студент 3 курса и как грамматические связи (ГС) влияют на запоминание. Для этого были составлены последовательности слов.

1. Последовательности несвязанных слов по 3, 4, 5, 6, 7 и 8 слов.

2. Группы из 2 связанных слов (например, радиусы атомов). Группы соединялись в последовательности по 3, 4 и 5 групп.

3. Группы из 3 связанных слов (например, алюминий активно взаимодействует), которые объединялись в последовательности по 2 и 3 группы и еще по 2 и 3 группы с одним несвязанным словом на конце.

4. Группы из 4 связанных слов (например, имеет место неограниченная смешиваемость). Составленные последовательности содержали 2 группы. В конце к ним добавлялось 0, 1 и 2 несвязанных слова.

5. Предложения различной длины (от 6 до 20 слов).

Все слова и словосочетания случайным образом были выбраны из учебников по разделам химии. Поэтому использовались слова как общеупотребительного, так и специального характера.

Обработка результатов. В каждой совокупности однородных последовательностей была найдена самая длинная, которую мог запомнить студент. По ней су-

дили о количестве осознанных связей. Например, испытуемый запомнил последовательность из 2 групп по 3 связанных слова. Таким образом, количество запомненных слов равно 6 и количество осознанных связей – 4. Всего было получено 100 точек с двумя координатами: количество запомненных слов и количество осознанных связей. Полученная зависимость оказалась линейной. С помощью программы Statgraphics Plus for Windows было получено следующее уравнение:

$$Y = 4{,}61 + 0{,}81X. \qquad (1)$$

Коэффициент корреляции равен 0,9. Следовательно, студент 3 курса может запомнить 4,6 несвязанных слова и каждая грамматическая связь позволяет запомнить на 0,8 слова больше. На основе этой модели вычисляем максимальное количество запоминаемых слов Y. Так как межсловесные связи не могут образовывать замкнутых контуров и связи с последним словом не содействуют первоначальному запоминанию, то количество осознанных связей во время чтения последнего слова не может превысить Y – 2. Подставляя данное выражение в наше уравнение, получаем Y = 15,8 слова. Таким образом, максимальное количество слов, которое студент 3 курса может запомнить с первого раза равно 15,8. Также выяснилось: а) короткие слова запоминаются лучше, чем длинные; б) повторы способствуют лучшему запоминанию.

Эксперимент 2. В предложении помимо грамматических существуют еще и семантические связи (СС), которые также влияют на запоминание. СС между словами определяются 2 факторами: а) совместное употребление этих слов; б) осознание связи между объектами действительности, которые обозначаются этими словами.

Очевидно, что СС между отдельными парами слов для разных людей различна, что зависит от подготовленности читателя, его уровня образования и т. д. Поэтому для того, чтобы определить силу СС для различных слов, был поставлен эксперимент с помощью метода словесных ассоциаций. Испытуемым зачитывали слово-стимул и они должны были записать первое пришедшее на ум слово. Общее количество слов-стимулов равнялось 140. О силе СС судили по количеству ответов на слово стимул. Например, на слово "кислота" 10 человек из 20 ответили серная и сила семантической связи между ними равна 10/20 = 0,5. Наибольшая возможная сила СС равна 1. Эксперименты, проведенные на 3 и 5 курсах, доказали тот факт, что сила СС между словами для отдельных людей различна. Например, на 3 курсе на слово-стимул "принцип" большинство ответов было Ле-Шателье, на 5 – действия.

Таким образом, был составлен список слов, в которых сила СС изменялась от 0,07 до 1. По этому списку были составлены следующие последовательности:

1. Последовательности по 3 и 4 пары, где в каждой паре сила СС не превышала 0,23.
2. Последовательности по 3, 4 и 5 пар (сила СС не превышала 0,46).
3. Последовательности по 3, 4 и 5 пар (сила СС не превышала 0,69).
4. Последовательности по 4 и 5 пар (сила СС была выше 0,69).

Обработка результатов. Данные эксперимента обрабатывались следующим образом: если в первом эксперименте все ГС считались равными, то во втором – силы СС не равны. Поэтому сила СС последовательности находилась как сумма сил СС групп, входящих в последовательность. Например, если студент запомнил последовательность – кислота, серный (0,42), сила, тяжесть (0,33), молекула, ве-

щество (0,39), то количество запомненных слов равно 6, а сила СС последовательности – 1,14. По результатам эксперимента с помощью программы Statgraphics Plus for Windows было получено следующее уравнение:

$$Y = 4{,}77 + 1{,}18X. \qquad (2)$$

Коэффициент корреляции равен 0,74. Следовательно, студент 3 курса может запомнить 4,77 несвязанных слова и каждая грамматическая связь позволяет запомнить на 1,18 слова больше.

Итак, по результатам экспериментов можно составить уравнение, с помощью которого можно определить трудность связей слова. Оно будет иметь следующий вид:

$$R = Y - 0{,}81X - 1{,}18X_1, \quad (3)$$

где R – трудность связей слова;

Y – количество слов, необходимых запомнить для образования самой длинной назадидущей связи этого слова;

X – количество уже образовавшихся грамматических связей между теми же словами;

X_1 – количество семантических связей между словами.

Таким образом, цель автора и редактора – ограничить трудность связи слова. Это позволит читателю осознавать связи с первого раза чтения и не возвращаться к уже прочитанному, теряя время.

Эксперименты показали, что объем оперативной памяти студентов лежит в пределах магического числа 7±2. По-видимому, эта цифра является универсальной для всех людей. Этот вывод наталкивает на мысль, что тогда и трудность связи слова в предложении не должна превышать этого предела, независимо от вида и уровня текста. Т. е. длина предложения не должна сильно различаться по видам текста (так как глубина фразы в большое степени зависит от ее длины [3]). Мы решили рассчитать среднюю длину предложения в следующих изданиях: учебниках по химии для 9 и 11 классов средней школы, учебниках по разделам химии для студентов вузов, научном журнале НАНБ (серия химических наук). Длина предложения рассчитывалась в словах.

Результаты подсчетов приведены ниже.

Средний размер простого (1), сложного (2) и всего предложения (3) в изданиях по химии

9-й класс			11-й класс			Учебники для вузов				Журнал НАНБ			
1	2	3	1	2	3		1	2	3		1	2	3
7,18	16,04	14,51	11,5	14,33	14,09	АХ	8,82	19,41	16,26	ВМС	13,61	23,95	18,87
						ОХ	8,83	20,17	16,40	АХ	13,83	24,48	18,98
						КХ	10,21	17,61	17,24	ОХ	15,39	25,17	19,23
						ФХ	12,72	23,21	17,28	ХТ	15,50	25,99	20,36
						НХ	13,79	23,6	17,40	КХ	15,66	27,15	20,55
										ГХ	16,22	27,74	20,67
										НХ	16,33	27,83	20,72
										ФХ	17,09	29,22	21,43
										БХ	17,30	30,14	22,07

Обозначения: ВМС – химия высокомолекулярных соединений; АХ – аналитическая химия; ОХ – органическая химия; ХТ – химические технологии; КХ – коллоидная химия; ГХ – геохимия; НХ – неорганическая химия; ФХ – физическая химия; БХ – биоорганическая химия.

Мы видим, что средняя длина простого, сложного и всего предложения увеличивается от 9-го класса к научному журналу НАНБ.

Эти данные согласуются с данными других исследователей [4-7], которые также показали, что средняя длина предложения для разных авторов, видов и жанров лежит в определенных интервалах. Исходя из этого, различные исследователи используют этот факт для различных целей. Например, Дж. Юл использовал в своем анализе размеры предложений для атрибуции спорного текста [4], Г.А. Лесскис говорит о зависимости между размером предложения и содержанием текста [5], а также о прямой зависимости между средней длиной предложения и количеством сложных предложений. Мы думаем, что по этим данным можно также судить о том, для какой группы читателей текст будет представлять наименьшую трудность. Например, после анализа химического текста средняя длина предложения составила 19,5. С большой долей вероятности можно утверждать, что он написан для читателя-специалиста (научного работника) и для студентов вузов он будет представлять определенную трудность. Правда, остается неясным, является ли эта длина оптимальной для данного читателя. Ведь основную трудность представляют превышающие предельную трудность межсловесные связи.

Анализ длины предложений научного журнала НАНБ показал, что порой предложения достигают 100 слов и выше. Но если редактор пропускает такое предложение, то, видимо, даже в нем межсловесные связи не превышают предела 7±2. Но, по всей вероятности, существуют и другие факторы, кроме трудностей межсловесных связей, которые затрудняют понимание длинных предложений.

Таким образом, цель редактора – не допустить превышающие предельную трудность межсловесные связи. Это позволит оптимизировать сложность текста и в максимальной степени удовлетворить информационные потребности читателей.

Литература

2. Миллер Дж. Магическое число 7±2. – В сб.: Инженерная психология. М., 1964.
3. Микк Я.А. Понятность учебного текста и связи в нем. – В кн.: Советская педагогика и школа. Тарту, 1970, вып. 2, С. 6-67.
4. Лущихина И.М. Использование гипотезы Ингве о структуре фразы при изучении восприятия речи // Вопросы психологии. – 1965. – № 2. – С. 57–67.
5. Yule G.U. On sentence-length as a statistical characteristic of style in prose: with application to two cases of disputed authorship, "Biometrica", XXX 3-4, 1939.
6. Лесскис Г.А. О зависимости между размером предложения и характером текста // Вопросы языкознания. – 1963. – № 3. – С. 92-112.
7. Лесскис Г.А. О зависимости между размером предложения и его структурой в разных видах текста // Вопросы языкознания. – 1964. – № 3. – С. 99-123.
8. Кравчук Н.В. Взаимодействие количественного и качественного аспектов в структуре предложения английского языка. Автореф. … канд. филол. наук. Мн., 1972.