

НАУЧНО • ТЕХНИЧЕСКАЯ ИНФОРМАЦИЯ

Серия 2. ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ И СИСТЕМЫ
ЕЖЕМЕСЯЧНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ СБОРНИК

Издается с 1961 г.

№ 3

Москва 2015

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ

УДК 004.89 : 510.64

Д.В. Виноградов

Вероятность порождения случайного ДСМ-сходства при наличии контр-примеров*

«Комбинаторный взрыв» числа ДСМ-гипотез можно объяснить возникновением ДСМ-сходств лишь из-за случайного совпадения нескольких несущественных признаков в обучающих примерах. Стандартными механизмами уменьшения числа гипотез являются увеличение границы на число примеров-родителей и запрет на контр-примеры. Рассматривая испытания Бернулли (для каждого обучающего примера и каждого атрибута) с вероятностью успеха p_j для j -го несущественного атрибута, получаем модель случайной обучающей выборки для возникновения случайного ДСМ-сходства. Таким же методом задаем множество контр-примеров. Для этой модели в настоящей статье вычисляется производящая функция возникновения случайного сходства с числом b примеров-родителей при t потенциальных контр-примерах.

Ключевые слова: испытания Бернулли, ДСМ-гипотеза, контр-пример, производящая функция, цепь Маркова

* Работа выполнена при поддержке программы фундаментальных исследований Президиума РАН «Математическое моделирование и интеллектуальные системы» на 2015 г. и гранта РФФИ (П8 и проект 14-07-00856а).

ВВЕДЕНИЕ

ДСМ-метод правдоподобных рассуждений является одним из лучших методов интеллектуального анализа данных (см. В.К Финн [1]). Он объединяет в себе индуктивное обобщение обучающих примеров в гипотезах о причинах, доопределение свойств неизвестных примеров по аналогии с обучающими примерами и абдуктивное принятие гипотез. Но теоретически возможна ситуация (Утверждение 1), когда число гипотез будет экспоненциально велико по сравнению с объемом обучающей выборки. В экспериментах довольно часто наблюдается «комбинаторный взрыв» числа гипотез. Чтобы справиться с такой ситуацией, ДСМ-метод предлагает два стандартных механизма: запрет гипотез, имеющих контр-примеры, и порождение гипотез из $\geq b$ примеров-родителей.

Каков же механизм появления такого огромного числа гипотез на реальных данных? Думается, что причина – появление несущественных сходств, возникающих из-за случайных совпадений множества атрибутов у нескольких обучающих примеров, каждый из которых имеет различные реальные причины. Простейший пример этого феномена приведен в следующем разделе (Пример 4). Такие «случайные» сходства могли бы быть уничтожены контр-примерами, но примеров противоположного знака обычно ограниченное число. Да и добавление их в базу фактов в значительной степени является делом случая.

Мы будем рассматривать модель формирования множества контр-примеров и обучающей выборки, основанную на независимых последовательностях испытаний Бернулли. Так как для возникновения «случайных» сходств необходимо появление общего подмножества несущественных (не входящих ни в какую реальную причину) признаков, то мы ограничимся рассмотрением только таких признаков. Всюду в настоящей работе число N обозначает их количество. Для каждого контр-примера или элемента обучающей выборки образуем последовательность испытаний Бернулли с вероятностью успеха p_j для j -го несущественного атрибута, причем последовательности для разных объектов – независимы. Число m будет равно числу контр-примеров. Число родителей будет равно b , но мы заменим (по Лемме 1) процедуру вычисления их сходства одной последовательностью испытаний Бернулли с вероятностью успеха $(p_j)^b$ для j -го несущественного атрибута.

Нас будет интересовать вероятность $P(X_N^{(m)} = 0)$ того, что при наличии m контр-примеров возникнет случайное сходство, которое не устранилось ни одним из этих контр-примеров. В настоящей статье вычисляется производящая функция $\sum_{m=0}^{\infty} P(X_N^{(m)} = 0) \cdot z^m$ и получаются формулы для $P(X_N^{(m)} = 0)$ при любых фиксированных b и N .

1. НЕОБХОДИМЫЕ ОПРЕДЕЛЕНИЯ И ПРИМЕРЫ

Определение 0. *Выборкой* обучающих примеров одного знака называется множество объектов O , каждый из которых описывается подмножеством множества признаков F . Отношение «объект $o \in O$ имеет признак $f \in F$ » задается $E(o, f) = 1$ или, что обычно для бинарных отношений, oEf . Его дополнение в $O \times F$ задается $E(o, f) = 0$.

Пример 1. Рассмотрим конечное множество объектов $O = \{o_1, \dots, o_n\}$, j -й из которых o_j описывается подмножеством $F^{-j} = \{f_1, \dots, f_{j-1}, f_{j+1}, \dots, f_n\}$ ($j = 1, \dots, n$) множества признаков $F = \{f_1, \dots, f_n\}$, непересекающегося с O . Случаю трех примеров соответствует таблица:

$O \setminus F$	f_1	f_2	f_3
o_1	0	1	1
o_2	1	0	1
o_3	1	1	0

Определение 1. *Пересечением* подмножества $A \subseteq O$ объектов назовем подмножество $A' = \{f \in F : \forall o \in A [oEf]\}$. Другими словами, пересечение $A' \subseteq F$ подмножества $A \subseteq O$ объектов – это все признаки, общие для всех объектов из множества A . Заметим, что пересечением пустого подмножества объектов является все множество признаков $\emptyset' = F$.

Пример 2. Результат операции пересечения подмножества $\{o_{j_1}, \dots, o_{j_k}\}$ объектов из примера 1 получается вычитанием из F подмножества признаков $\{f_{j_1}, \dots, f_{j_k}\}$.

Определение 2. *Пересечением* подмножества $B \subseteq F$ признаков назовем подмножество $B' = \{o \in O : \forall f \in B [oEf]\}$. Другими словами, пересечение $B' \subseteq O$ множества $B \subseteq F$ признаков – это все объекты, обладающие всеми признаками из множества B . Пересечением пустого подмножества признаков является все множество объектов $\emptyset' = O$.

Пример 3. Результат операции пересечения подмножества $\{f_{j_1}, \dots, f_{j_k}\}$ признаков из Примера 1 получается вычитанием из O подмножества объектов $\{o_{j_1}, \dots, o_{j_k}\}$.

Определение 3. *ДСМ-сходством* из множества объектов O , каждый из которых описывается подмножеством множества признаков F называется всякая пара (A, B) , состоящая из подмножества (называемого *списком родителей*) объектов $A \subseteq O$ и

подмножества признаков $B \subseteq F$ (называемого **пересечением объектов**), удовлетворяющая условиям $A' = B$ и $B' = A$ (последнее условие называется **исчерпываемостью**). ДСМ-сходства ограничивают дополнительным требованием $|A| \geq b$, где b - фиксированный порог для минимального числа примеров-родителей. В классическом определении $b = 2$.

Утверждение 1. *Существует обучающая выборка для ДСМ-метода с n объектами одного знака, для которой возможно возникновение*

$$2^n - \sum_{j=0}^{b-1} \binom{n}{j} - 1 \text{ ДСМ-сходств с не менее, чем } b \text{ родителями.}$$

Доказательство. Достаточно рассмотреть При-

мер 1. Согласно примеру 2 все 2^n подмножеств будут пересечениями соответствующих подмножеств объектов. Но ДСМ-метод исключит из списка сходств

$$1 = \binom{n}{0} \text{ пересечение пустого подмножества, } n = \binom{n}{1}$$

$$\text{одноэлементных подмножеств объектов, } \dots, \binom{n}{b-1}$$

$$(b-1)\text{-элементных подмножеств объектов и } 1 = \binom{n}{n}$$

пересечение всех объектов, так как оно будет иметь пустое подмножество признаков, согласно Примеру 3.

Пример 4. Рассмотрим множество самолетов $O = \{o_1 = B-737, SSJ-100, IL-76, A-320\}$, находящихся на ремонте, каждый из которых описывается проблемами из списка

$$F = \{f_1 = \text{wing}, f_2 = \text{engine}, f_3 = \text{blasphemy}\}$$

$O \setminus F$	f_1	f_2	f_3
o_1	1	0	0
o_2	1	0	1
o_3	0	1	1
o_4	0	1	0

Перечислим все ДСМ-сходства в таблице:

i	A	B
0	$\emptyset$	$\{f_1, f_2, f_3\}$
1	$\{o_2\}$	$\{f_1, f_3\}$
2	$\{o_3\}$	$\{f_2, f_3\}$
3	$\{o_1, o_2\}$	$\{f_1\}$
4	$\{o_2, o_3\}$	$\{f_3\}$
5	$\{o_3, o_4\}$	$\{f_2\}$
6	$\{o_1, o_2, o_3\}$	$\emptyset$

При $b = 2$ получаются две «настоящие» гипотезы ($\{o_1, o_2\}, \{f_1\}$) «Самолет с поврежденным крылом не летает» и ($\{o_3, o_4\}, \{f_2\}$) «Самолет с поврежденным двигателем не летает» и одна «случайная» ($\{o_2, o_3\}, \{f_3\}$) «Самолет, на котором написано ругательство, не летает». Эта последняя гипотеза возникла из-за случайного совпадения подмножества признаков $\{f_3\}$ у двух примеров o_2 и o_3 , каждый из которых успешно объясняется различными «настоящими» гипотезами. Не существует никакого очевидного способа отличить «случайную» гипотезу от «настоящей».

Определение 4. *Последовательность испытаний Бернулли* – это распределение вероятностей на

$$\{0, 1\}^N \text{ с } \mathbf{P}(f_1 = \sigma_1, \dots, f_N = \sigma_N) = \prod_{j=1}^N p_j^{\sigma_j} \cdot (1 - p_j)^{1 - \sigma_j},$$

где $\sigma_j \in \{0, 1\}$ и $0 < p_j < 1$. Число $0 < p_j < 1$ называется вероятностью успеха $f_j = 1$ в j -ом испытании.

В настоящей статье контр-примеры и элементы обучающей выборки соответствуют независимым реализациям элементов этого вероятностного пространства.

Лемма 1. *Случайное сходство $(a_1, \dots, a_N) \in \{0, 1\}^N$ b обучающих примеров является последовательностью испытаний Бернулли с вероятностью успеха $a_j = 1$, равной $(p_j)^b$.*

Доказательство следует из условия независимости обучающих примеров и Определения 1.

Определение 5. Контр-пример $(f_1^j, \dots, f_N^j) \in \{0, 1\}^N$ ($j = 1, \dots, m$) *устраняет* (или *устраняется*) случайное(-ым) сходство(-м) $(a_1, \dots, a_N) \in \{0, 1\}^N$, если $\forall k [a_k = 1 \Rightarrow f_k^j = 1]$.

2. ЦЕПЬ МАРКОВА ДЛЯ УСТРАНЕНИЯ КОНТР-ПРИМЕРОВ

Определение 6. *(Конечной) цепью Маркова* на множестве состояний $S = \{0, \dots, m\}$ назовем $(m+1) \times (m+1)$ -матрицу $\mathbf{P} = (p_{i,j})$ из неотрицательных действительных чисел, удовлетворяющую условию, что все суммы по строкам равны 1: $\forall i [\sum p_{i,j} = 1]$. Матрицы такого вида называются **стохастическими**. Элемент $p_{i,j}$ этой матрицы – вероятность перехода в момент времени $t+1$ в состояние $X_{t+1} = j$, если в момент времени t стартуем из состояния $X_t = i$.

Пример 5. Рассмотрим процесс одновременного нахождения $t+1$ -х признаков сходства a_{t+1} и всех контр-примеров f_{t+1}^j ($j = 1, \dots, m$). В момент времени t состоянием $X_t^{(m)}$ цепи Маркова будет число неуст-

раненных после одновременного нахождения t -х признаков контр-примеров. Если первоначально было m контр-примеров (на это указывает верхний индекс), то $X_0^{(m)} = m$ с вероятностью 1. Нас интересует $\mathbf{P}(X_N^{(m)} = 0)$ – вероятность того, что после определения всех N признаков, все m контр-примеров будут устранены.

Лемма 2. Для примера 1 матрица перехода равна

$$p_{s+r,s} = \mathbf{P}(X_{t+1}^{(m)} = s | X_t^{(m)} = s+r) = \begin{cases} (1 - (p_{t+1})^b) + (p_{t+1})^b \cdot p_{t+1}^m, & r = 0 \\ (p_{t+1})^b \cdot \binom{s+r}{r} \cdot (1 - (p_{t+1}))^r \cdot (p_{t+1})^s, & 0 < r \leq m-s \end{cases}.$$

Доказательство. Если после определения t -х признаков оставалось $s+r$ ($0 < r \leq m-s$) неустраненных контр-примеров, а после определения $t+1$ -х их осталось s , то $t+1$ -й признак $a_{t+1} = 1$ и какие-то r из $s+r$ уцелевших контр-примеров должны получить 0 в $t+1$ -й позиции, а остальные s должны получить 1 в $t+1$ -й позиции. Верхняя строчка вычисляется разбором случаев, когда $t+1$ -й признак $a_{t+1} = 0$, и когда $a_{t+1} = 1$ и все $s+r$ уцелевших контр-примеров получают 1 в $t+1$ -й позиции.

Определение 7. Производящей функцией для распределения вероятностей $\{w_s \geq 0: s \in S\}$ (где $\sum_{s=0}^m w_s = 1$) на множестве состояний $S = \{0, \dots, m\}$ на-

зовем многочлен $\varphi(x) = \sum_{s=0}^m w_s \cdot x^s$.

Пример 6. Производящие функции для распределений $\mathbf{P}(X_t^{(m)} = s)$ обозначаются $\varphi_t^{(m)}(x) = \sum_{s=0}^m \mathbf{P}(X_t^{(m)} = s) \cdot x^s$. Очевидно, что $\varphi_0^{(m)}(x) = x^m$. При этом нас интересует число $\mathbf{P}(X_N^{(m)} = 0) = \varphi_N^{(m)}(0)$.

Лемма 3. Для примера 6 производящие функции преобразуются так:

$$\varphi_{t+1}^{(m)}(x) = (1 - p_{t+1}^b) \cdot \varphi_t^{(m)}(x) + p_{t+1}^b \cdot \varphi_t^{(m)}(p_{t+1} \cdot x + (1 - p_{t+1})).$$

Доказательство.

$$\text{Пусть } \varphi_t^{(m)}(x) = \sum_{s=0}^m w_s \cdot x^s \text{ и } \varphi_{t+1}^{(m)}(x) = \sum_{s=0}^m v_s \cdot x^s,$$

$$\text{где } w_s = \mathbf{P}(X_t^{(m)} = s) \text{ и } v_s = \mathbf{P}(X_{t+1}^{(m)} = s).$$

Тогда

$$v_s = \mathbf{P}(X_{t+1}^{(m)} = s) = \sum_{r=0}^{m-s} \mathbf{P}(X_{t+1}^{(m)} = s | X_t^{(m)} = s+r) \cdot \mathbf{P}(X_t^{(m)} = s+r) =$$

$$\sum_{r=0}^{m-s} \mathbf{P}(X_{t+1}^{(m)} = s | X_t^{(m)} = s+r) \cdot w_{s+r} = (1 - p_{t+1}^b) \cdot w_s + \sum_{r=0}^{m-s} p_{t+1}^b \cdot \binom{s+r}{r} \cdot (1 - p_{t+1})^r \cdot p_{t+1}^s \cdot w_{s+r}$$

(последнее равенство следует из Леммы 2).

Поэтому

$$\begin{aligned} \varphi_{t+1}^{(m)}(x) &= \sum_{s=0}^m v_s \cdot x^s = (1 - p_{t+1}^b) \cdot \sum_{s=0}^m w_s \cdot x^s + \\ &+ p_{t+1}^b \cdot \sum_{s=0}^m \sum_{r=0}^{m-s} \binom{s+r}{r} \cdot (1 - p_{t+1})^r \cdot w_{s+r} \cdot (p_{t+1} \cdot x)^s = \\ &= (1 - p_{t+1}^b) \cdot \varphi_t^{(m)}(x) + p_{t+1}^b \cdot \sum_{k=0}^m \sum_{r=0}^k \binom{k}{r} \cdot \\ &\cdot (1 - p_{t+1})^r \cdot w_k \cdot (p_{t+1} \cdot x)^{k-r} = \end{aligned}$$

(где ввели обозначение $k = s+r$)

$$= (1 - p_{t+1}^b) \cdot \varphi_t^{(m)}(x) + p_{t+1}^b \cdot \sum_{k=0}^m w_k \cdot \left(\sum_{r=0}^k \binom{k}{r} \cdot (1 - p_{t+1})^r \cdot (p_{t+1} \cdot x)^{k-r} \right) =$$

(по формуле бинома Ньютона)

$$\begin{aligned} &= (1 - p_{t+1}^b) \cdot \varphi_t^{(m)}(x) + p_{t+1}^b \cdot \sum_{k=0}^m w_k \cdot (p_{t+1} \cdot x + (1 - p_{t+1}))^k = \\ &= (1 - p_{t+1}^b) \cdot \varphi_t^{(m)}(x) + p_{t+1}^b \cdot \varphi_t^{(m)}(p_{t+1} \cdot x + (1 - p_{t+1})). \end{aligned}$$

Замечание 1. Преобразование $T_t q(x) = (1 - p_t^b) \cdot q(x) + p_t^b \cdot q(p_t \cdot x + (1 - p_t))$ задает линейный оператор на многочленах. При этом этот оператор не зависит от m . Согласно рассуждениям в Примерах 5 и 6 нас интересует $\mathbf{P}(X_N^{(m)} = 0) = \varphi_N^{(m)}(0)$, где $\varphi_N^{(m)}(x) = \left(\prod_{t=1}^N T_t \right) x^m$, так как $\varphi_0^{(m)}(x) = x^m$.

Определение 8. Собственным базисом линейного оператора $T: L \rightarrow L$, действующего в векторном пространстве L называется базис из собственных векторов.

Лемма 4. Для операторов T_t из Замечания 1 собственный базис образуют многочлены $(x-1)^j$ ($j = 0, 1, 2, \dots$) с собственными значениями $p_t^b \cdot p_t^j + (1 - p_t^b)$.

Доказательством является непосредственная проверка.

$$\text{Лемма 5. } \varphi_0^{(m)}(x) = \sum_{j=0}^m \binom{m}{j} \cdot (x-1)^j - \text{разложение}$$

по собственному базису.

$$\text{Доказательство. По биному Ньютона } \varphi_0^{(m)}(x) = x^m = (1 + (x-1))^m = \sum_{j=0}^m \binom{m}{j} \cdot (x-1)^j.$$

Теорема 1.

$$P(X_N^{(m)} = 0) = \sum_{j=0}^m \binom{m}{j} \cdot (-1)^j \cdot \prod_{t=1}^N [p_t^b \cdot p_t^j + (1 - p_t^b)].$$

Доказательство. Собираем вместе утверждения Замечания 1 и Лемм 4 и 5.

Замечание 2. Результат Теоремы 1 был получен автором ранее помощью принципа «включение-исключение».

3. ПРОИЗВОДЯЩАЯ ФУНКЦИЯ ВЕРОЯТНОСТЕЙ СЛУЧАЙНОГО СХОДСТВА В ЗАВИСИМОСТИ ОТ ЧИСЛА КОНТР- ПРИМЕРОВ

Определение 9. Производящей функцией для последовательности вероятностей $P(X_N^{(m)} = 0)$ (где $m = 0, 1, 2, \dots$) назовем формальный степенной ряд $\varphi_N(0, z) = \sum_{m=0}^{\infty} P(X_N^{(m)} = 0) \cdot z^m$. Двойной производящей функцией для вероятностей $P(X_N^{(m)} = s)$ назовем формальный ряд

$$\varphi_N(x, z) = \sum_{m=0}^{\infty} \sum_{s=0}^m P(X_N^{(m)} = s) \cdot x^s \cdot z^m = \sum_{m=0}^{\infty} \varphi_N^{(m)}(x) \cdot z^m.$$

Лемма 6. $\varphi_0(x, z) = \sum_{j=0}^{\infty} \frac{z^j}{(1-z)^{j+1}} \cdot (x-1)^j$ – разложение по собственному базису.

Доказательство.

$$\begin{aligned} \varphi_0(x, z) &= \sum_{m=0}^{\infty} x^m \cdot z^m = \frac{1}{1-z \cdot x} = \frac{1}{1-z} \cdot \frac{1}{1 - \frac{z \cdot (x-1)}{1-z}} = \\ &= \sum_{j=0}^{\infty} \frac{z^j}{(1-z)^{j+1}} \cdot (x-1)^j \end{aligned}$$

Теорема 2.

$$\varphi_N(0, z) = \sum_{j=0}^{\infty} \prod_{t=1}^N [p_t^b \cdot p_t^j + (1 - p_t^b)] \cdot \frac{(-z)^j}{(1-z)^{j+1}}.$$

Доказательство аналогично доказательству Теоремы 1, но вместо Леммы 5 нужно использовать Лемму 6 совместно с Леммой 4.

ЗАКЛЮЧЕНИЕ

В настоящей работе получены точные формулы для вероятности $P(X_N^{(m)} = 0)$ того, что при наличии m контр-примеров возникнет случайное сходство, которое не устранится ни одним из этих контр-примеров. Эту формулу можно использовать как оценку среднего числа «случайных гипотез», так как по обучающей выборке и множеству контр-примеров можно оценить вероятность появления каждого признака (как случайную величину Бернулли), а число признаков и контр-примеров известно.

Выведенный в заключительном параграфе функциональный ряд производящей функции для последовательности вероятностей $P(X_N^{(m)} = 0)$ (где $m = 0, 1, 2, \dots$) может привести к асимптотическим оценкам этих вероятностей. Можно доказать, что полученный в Теореме 2 функциональный ряд сходится при $(z < 1/2)$. Автор призывает специалистов по асимптотическим методам исследовать приведенный ряд хотя бы в простейшем случае, когда все вероятности появления признаков p_t равны друг другу.

Изложенные в работе результаты в равной степени применимы как к ДСМ-методу [1] автоматического порождения гипотез, так и к Анализ Формальных Понятий (АФП) (расширенному добавлением контр-примеров) [2], и к ВКФ-методу, развиваемому автором [3] со своими студентами и аспирантами. Более того, автор полагает, что именно из-за неизбежного появления случайных сходств исчерпывающая процедура вычисления всех сходств, лежащая в основе ДСМ-метода и АФП, приводит к переобучению. ВКФ-метод имеет более гибкие механизмы управления множеством порожденных гипотез и может ограничиваться из-за своего вероятностного характера более робастными моделями.

* * *

Автор благодарит М.А. Михеенкову за поддержку и Т.А. Волкову за полезные обсуждения.

СПИСОК ЛИТЕРАТУРЫ

1. Финн В.К. О машинно-ориентированной формализации правдоподобных рассуждений в стиле Ф. Бэкона – Д. С. Милля // Семиотика и информатика. – 1983. – Вып. 20. – С. 35-101.
2. Ganter B., Kuznetsov S.O. Formalizing Hypotheses with Concepts. // Proc. 8th International Conference on Conceptual Structures ICCS 2001 / eds. G. Mineau, B. Ganter. – Lecture Notes in Artificial Intelligence (Springer). – 2000. – Vol. 1867. – P. 342-356.
3. Vinogradov D.V. VKF-method of Hypotheses Generation // Analysis of Images, Social Networks and Texts / eds. D.I. Ignatov et al. – Communications in Computer and Information Science (Springer). – 2014. – Vol. 436. – P. 237-248.

Материал поступил в редакцию 23.12.14.

Сведения об авторе

ВИНОГРАДОВ Дмитрий Вячеславович – кандидат физико-математических наук, доцент РГТУ, старший научный сотрудник ВИНТИ РАН, Москва
e-mail: Vin@viniti.ru

Н.В. Харчевникова, В.Г. Блинова, Д.А. Добрынин, В.С. Журков

Использование интеллектуальной ДСМ-системы для анализа связи структуры нитрозамещенных бензолов с их мутагенной активностью в тесте Эймса*

Проведен анализ зависимости мутагенной активности в тесте Salmonella/микросомы на штаммах TA100 и TA98 в вариантах с активацией и без активации для соединений ряда замещенных нитробензолов. Обучающая выборка содержала данные по мутагенности (+/-) для 254 соединений. Анализ проведен с использованием интеллектуальной ДСМ-системы автоматического порождения гипотез, реализующей метод ДСМ, и языка представления данных, называемого фрагментарный код суперпозиции подструктур (ФКСП). Для проверки правильности прогноза проведен скользящий контроль. Причины ошибочных предсказаний изучены с использованием квантово-химических расчетных параметров.

Ключевые слова: интеллектуальная ДСМ-система, мутагенная активность, тест Эймса, нитробензол, язык представления данных (ФКСП), квантово-химические расчеты

ВВЕДЕНИЕ

Среди известных методов тестирования мутагенной активности химических соединений тест *Salmonella*/микросомы (тест Эймса) получил наибольшее распространение. За годы, прошедшие со времени появления теста Эймса (70-е гг. прошлого века) накоплены результаты применения этого теста на различных штаммах бактерий.

Штаммы TA100 (выявляющий мутации типа замены пар оснований) и TA98 (выявляющий мутации типа сдвига рамки считывания) рекомендованы в качестве индикаторов мутагенной активности ксенобиотиков. Для более полной имитации действия ксенобиотиков на организм млекопитающих разработан вариант теста с добавлением постмитохондриального супернатанта гомогената печени (S9mix фракция) крыс или других грызунов (тест с биоактивацией) [1]. Компоненты S9mix содержат целые комплексы ферментов биотрансформации химических соединений. В мире ежегодно появляются сотни новых веществ, мутагенная активность которых не изучена. Результаты применения теста Эймса для большого числа соединений, содержатся в базах данных TOXNET и ISSSTY [2, 3], что позволяет проводить *in silico* прогноз мутагенной активности веществ. Известно, что в отсутствие S9mix могут происходить реакции восстановления нитрогруппы, реакции с тиоловыми группами биомолекул, гидролиз. В присутствии S9mix может происходить окисление аминогруппы, гидроксильро-

вание бензольного кольца, реакции алифатического окисления групп-заместителей в бензольном кольце. Сложность состоит в том, что очень близкие по структуре соединения (изомеры) часто либо обладают мутагенностью, либо она отсутствует.

Электрон-акцепторная нитрогруппа препятствует протеканию реакции гидроксильирования бензольного кольца [4]. Поэтому, основным путем метаболической активации молекул большинства нитрозамещенных ароматических соединений является восстановление нитрогруппы под действием нитроредуктаз.

Различают кислород-зависимые и кислород-независимые нитроредуктазы [5-6]. Бактериальные нитроредуктазы относятся к кислород-независимым. При действии таких нитроредуктаз происходит перенос на субстрат двух электронов, т.е. перенос гидрид-иона [7]. В случае динитро- и тринитрозамещенных бензолов возможно присоединение гидрид-иона к атому углерода бензольного кольца, в частности к атому, замещенному нитрогруппой, с образованием на первой стадии интермедиатов, содержащих тетраэдрически координированный атом углерода бензольного кольца, с последующим отрывом нитрогруппы. Возможно также присоединение гидрид-иона к атому кислорода нитрогруппы с последующим образованием нитрозобензолов и чрезвычайно быстрым образованием гидроксиламинов, которые могут далее восстанавливаться до аминов с образованием интермедиатов – электрофильных нитрениевых ионов. Нитрениевые ионы способны образовывать аддукты с ДНК [4].

В клетках бактерий *S. typhimurium* идентифицированы две кислород-независимые нитроредуктазы: *SnrA* and *cnr*. Они принадлежат к типу осуществ-

* Работа выполнена при поддержке гранта РФФИ (проект 14-07-00856а) и программы фундаментальных исследований Президиума РАН (П8, проект №209)

ляющих активацию нитросоединений до мутагенов, определяющих мутагенность в тесте Эймса. [8-9]. Штаммы, используемые в тесте Эймса, не изогенны по отношению к присутствию *SnrA* и *spn* генов, что определяет субстратную специфичность этих штаммов, т.е. различие результатов теста Эймса одного и того же соединения на различных штаммах *S.typhimurium*.

Количественное соотношение структура-мутагенная активность в ряду нитрозамещенных ароматических соединений в тесте Эймса изучено в работе [10]. Однако обучающая выборка содержала полиароматические и гетероциклические соединения, что определило зависимость мутагенности в основном от параметра *lgP*, характеризующего гидрофобность соединений. Известно, что этот параметр сильно зависит от размера молекул, в частности от размера ароматической системы [10].

Компьютерная система OASIS [11, 12] содержит модуль для прогноза мутагенной активности в тесте Эймса на штаммах TA100 и TA98 без активации и с активацией. С использованием гетерогенной обучающей выборки, содержащей 2844 соединений, идентифицированы структурные требования для проявления мутагенности. Однако это коммерческая система, что затрудняет проверку ее применимости для прогноза мутагенности близких по структуре изомеров нитрозамещенных бензолов. Известно, что именно анализ данных по мутагенности изомеров замещенных нитробензолов в тесте Эймса является сложной задачей [13].

В 90-х гг. прошлого века экспертным сообществом были выделены функциональные группы и субструктуры, наличие которых в молекулах соединений определяет высокую вероятность проявления этими соединениями канцерогенной и мутагенной активности, поскольку эти группы либо сами могут вступать в реакции с ДНК в качестве электрофильных агентов, либо могут превращаться в электрофильные метаболиты [14-15]. Такие функциональные группы и субструктуры называются структурными «тревожными сигналами» (structural alerts).

Целью нашей работы является анализ мутагенной активности замещенных нитробензолов в тесте Эймса на штаммах TA100 и TA98 в вариантах с активацией и без активации с использованием интеллектуальной ДСМ-системы [16, 17], которая позволяет выявлять структурные фрагменты, характерные для активных и неактивных соединений. Результаты сопоставлены с результатами экспертной оценки, проведенной с использованием квантово-химических расчетных параметров, характеризующих процессы биоактивации и взаимодействия с ДНК самих нитробензолов или их метаболитов, а также со структурными «тревожными сигналами», выделенными ранее другими авторами.

МАТЕРИАЛЫ И МЕТОДЫ

Использованная в работе интеллектуальная ДСМ-система для исследования связи «структура химического соединения-биологическая активность» может быть представлена тремя уровнями: сходства, правил правдоподобного вывода и рассуждений. На первом уровне работы ДСМ-системы – уровне сходства – за-

дается операция сходства (выделение фрагмента структуры, общего для двух или более соединений, обладающих или не обладающих активностью). На втором уровне – уровне правил – соединения подразделяются на положительные (активные соединения) и отрицательные (неактивные соединения) примеры. С помощью определенных операций пересечения множеств находят сходства положительных примеров и сходства отрицательных примеров. Таким образом строятся гипотезы (фрагменты химической структуры) о причинах наличия или отсутствия данного свойства. На третьем уровне – уровне рассуждений – с помощью построенных гипотез осуществляется прогноз биологической активности химических соединений, информация об активности которых отсутствует.

Для прогноза мутагенности использованы два варианта ДСМ-метода: с запретом на контрпримеры и без запрета (простой метод). Запрет на контрпримеры означает, что фрагменты структуры, ответственные за наличие активности, и фрагменты, ответственные за ее отсутствие, не должны присутствовать в молекулах неактивных (положительные примеры) и активных (отрицательные примеры) соединений одновременно.

Обучающая выборка содержала результаты теста Эймса в альтернативной форме (+/-) для 254 нитрозамещенных бензолов. Компьютерные эксперименты проводились в нескольких вариантах. В настоящей работе обсуждаются результаты одного из них. Этот вариант соответствует выделению соединений, не проявляющих активности ни в одном тесте, ни на одном штамме. Таких соединений в выборке было 62. Остальные соединения считались положительными примерами.

Поскольку мутагенная активность нитрозамещенных бензолов связана с восстановлением нитрогруппы, а способность к восстановлению характеризуется значением нижней свободной молекулярной орбитали, были проведены расчеты электронного строения соединений выборки квантово-химическими методами AM1 и DFT.

Мутагенная активность нитрозамещенных бензолов может зависеть от их гидрофобности. Поскольку в выборке присутствовали соединения, которые при условиях проведения теста Эймса (нейтральная среда), находятся в ионизированном состоянии, для характеристики гидрофобности использовали также *lgD*, т.е. логарифм коэффициента распределения октанол/вода при нейтральном pH, равном 7.4. Для того чтобы соединение проявляло активность хотя бы в одном варианте теста Эймса, значения логарифмов коэффициентов распределения октанол/вода должны находиться в интервалах $-1 < \lg P < 6$; $-4,3 < \lg D < 4,5$.

Обсуждение результатов

Для проверки правильности прогноза в компьютерном эксперименте с отделением полностью неактивных соединений от соединений, показавших мутагенную активность хотя бы в одном варианте теста Эймса, был проведен скользящий контроль. Скользящий контроль состоит в последовательном исклю-

чении по одному каждого соединения из обучающей выборки и прогнозе для этого соединения как тестового. Применение метода ДСМ с запретом на контр-примеры дало возможность прогноза для 40% соединений в выборке. В том случае, когда метод с запретом на контрпримеры не позволял сделать прогноз, использовались результаты простого метода. Применение простого метода без запрета на контр-примеры позволило увеличить полноту прогноза до 75%. В результате такого подхода точность прогноза составила 85%, однако специфичность, т.е. процент правильно предсказанных отрицательных примеров была очень низкой. Результаты компьютерных ДСМ-экспериментов приведены в табл. 1.

Среди ошибочных предсказаний наиболее опасны те, которые предсказывают отсутствие мутагенной активности для мутагенных соединений. Таких «ошибочно отрицательных» предсказаний при скользящем контроле было сделано всего три: 4-нитробензилбромид (I), 4-амино-3-нитро-6-изопропиланилин (II), 4-нитробензилпропио-нат (III).

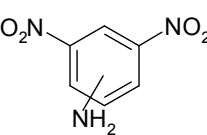
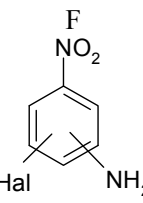
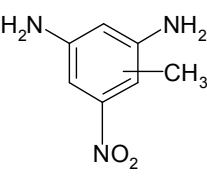
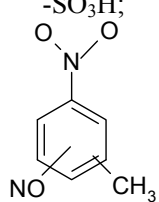
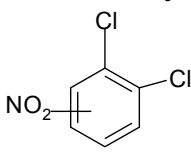
Мутагенность замещенных бензолов, содержащих группу CH_2Hal (где Hal-атом галогена), определяется возможностью формирования аддуктов с ДНК по механизму замещения $\text{S}_{\text{N}}1$.

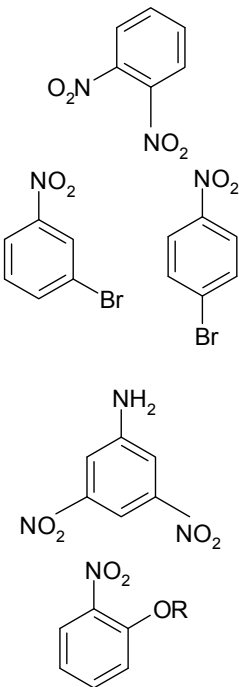
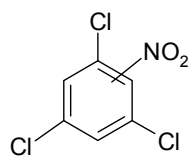
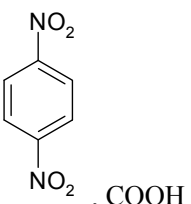
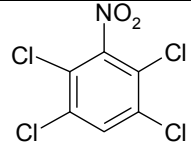
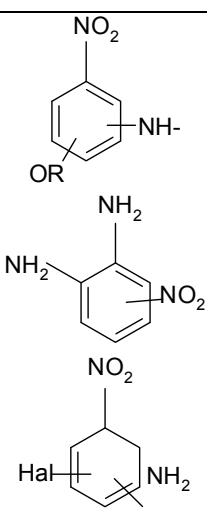
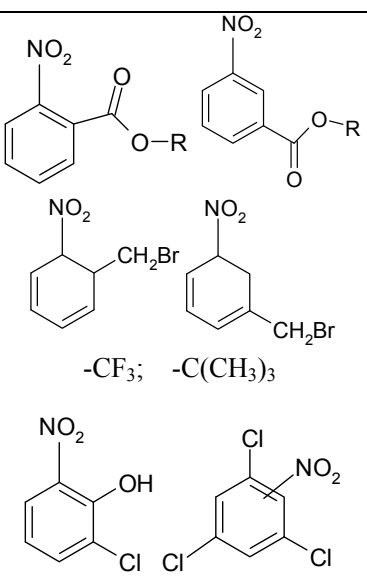
Орто- и *пара*-изомеры монозамещенных нитробензолов, содержащие группу CH_2Cl , проявляют мутагенную активность на штамме TA98, *пара*-изомер

активен также на штамме TA100. Однако *мета* и *орто*-изомеры CH_2Br замещенных нитробензолов не являются мутагенами в тесте Эймса, *пара*-изомер проявляет активность только на штамме TA100 без активации. Это соединение было предсказано как не обладающее активностью. Отрицательная гипотеза содержала фрагмент CH_2Br , и родители были *орто*- и *пара*-нитробензилбромиды. На основе этой гипотезы мы сформировали «категорию» структурно родственных соединений, объединенных одним механизмом действия, включив в нее нитробензолы, замещенные группой CH_2Cl . Различие мутагенности бромметил- и хлорметилзамещенных нитробензолов может объясняться слишком высокой стабильностью карбокатионов, образующихся при отрыве иона брома по сравнению с карбокатионами, образующимися при отрыве ионов хлора (бром является лучшей, чем хлор, уходящей группой). Эта разность определяет способность карбокатионов связываться с клеточными протеинами и повреждать клеточные структуры, т.е. цитотоксичность, мешающую проявлению мутагенности. В табл. 2 приведены рассчитанные методом AM1 теплоты образования галометилнитробензолов (ΔHf) и соответствующих карбокатионов ($\Delta\text{Hf}_{\text{cc}}$). Результаты теста Эймса представлены знаками +/- в порядке: TA100 без активации, TA100 с активацией, TA98 без активации, TA98 с активацией.

Таблица 1

Результаты компьютерных ДСМ-экспериментов

Вариант компьютерного эксперимента	Положительные гипотезы	Отрицательные гипотезы
1	2	3
1. Вариант с выделением полностью неактивных соединений и активных хотя бы в одном виде теста	$-\text{CH}_2\text{Cl};$ $-\text{COCl}; -\text{CH}=\text{CH}_2;$  $-\text{NH}-\text{NH}_2$ $-\text{COOH}$  	$-\text{SO}_3\text{H};$  

Вариант компьютерного эксперимента	Положительные гипотезы	Отрицательные гипотезы
1	2	3
2. Выделение соединений, мутагенных только на штамме TA100 с активацией и без активации (положительные примеры) и мутагенных только на штамме TA98 с активацией и без активации (отрицательные примеры)		-COCH_3 
3. Выделение соединений, мутагенных только на штамме TA98 (с активацией и без активации)		
4. Вариант с выделением соединений, активных во всех видах теста и полностью неактивных		

**Устойчивость карбокатионных интермедиатов галометилнитробензолов
и мутагенная активность на штаммах TA100 (без активации и с активацией)
и TA98 (без активации и с активацией)**

Название	ΔH_f ккал/моль	$\Delta H_{f_{cc}}$ ккал/моль	$\Delta \Delta H_{f_{cc}}$ ккал/моль	Мутагенность
п-(CH ₂ Br)нитробензол	23,14	243,9	220,76	[+--]
м-(CH ₂ Br)нитробензол	23,27	240,34	217,07	[----]
о-(CH ₂ Br)нитробензол	26,36	243,47	217,11	[----]
п-(CH ₂ Cl)нитробензол	10,87	243,9	233,03	[++++]
о-(CH ₂ Cl) нитробензол	13,90	243,47	229,57	[++++]

Видно, что карбокатионы, образующиеся при реакции замещения группы CH₂Br более устойчивы, чем в случае реакции CH₂Cl.

Соединение 4-амино-3-нитро-6-изопропиланилин (II) не обладает мутагенной активностью, тогда как 4-амино-3-нитро-5-изопропиланилин является мутагеном на штамме TA100, но только после активации с S9mix. Оценка энергий граничных орбиталей, рассчитанных методом DFT свидетельствует, что нитрогруппы в этих соединениях не должны подвергаться одноэлектронному восстановлению. Причиной мутагенности одного из двух родителей единственной отрицательной гипотезы (наличия изопропильной группы и замещенного группами NH- и NO₂ бензольного кольца), 4-амино-3-нитро-5-изопропиланилина, является большая устойчивость (по результатам расчета) радикала, образующегося при отрыве атома водорода от α -углеродного атома изопропильной группы под действием цитохрома P450. В молекуле второго родителя отрицательной гипотезы 2-метил-N-[4-нитро-3-(трифторметил)фе-нил]пропанамида изопропильная группа в отличие от тестового соединения не связана с бензольным кольцом.

пара-Нитробензилпропионат (III) является мутагеном в тесте Эймса, но был предсказан как не обладающий активностью. Все орто- и мета-нитробензиловые эфиры – ацетаты, пропионаты, бутираты не обладают мутагенной активностью, тогда как пара-изомеры мутагенны на штамме TA100 без активации. Различие мутагенности пара-изомеров от орто- и мета-изомеров возможно связано с особенностями связывания пара-изомеров с нитроредуктазами. Скорости возможного гидролиза изомеров мало отличаются [13].

Прогноз мутагенной активности неизученных соединений

Были проведены эксперименты по прогнозу мутагенности 16 соединений, входящих в обучающую выборку. Эти соединения были представлены как тест-примеры. Компьютерный эксперимент проводили по варианту выделения отрицательных примеров (вариант 1 в табл. 1). Соединения, показавшие активность хотя бы в одном варианте теста хотя бы на одном штамме, считались «положительными» примерами. В результате удалось спрогнозировать активность пяти соединений. Мутагенность всех этих соедине-

ний была предсказана верно. Затем был сделан прогноз для тринадцати соединений, не вошедших в обучающую выборку. Для девяти из них удалось сделать прогноз.

Результаты прогноза мутагенной активности неизученных соединений

- 6-бром-1,3-динитробензол (+)
- 2,4,6-триметилнитробензол (+)
- 3,5-дихлорнитробензол (+)
- 6-иод-1,3-динитробензол не определен
- 2,5-дибромнитробензол (+)
- 1,2-дихлор-4,5-динитробензол (+)
- 3-метил-4-бромнитробензол (не определен)
- 4,6-дихлор-1,2-динитробензол (+)
- 3,5-динитробензиловый спирт (не определен)
- 3,4-динитробензиловый спирт (+)
- 2,4,6-трихлор-1,3-динитробензол (-)
- 2,3,5,6-тетрахлор-1,4-динитробензол (не определен)
- 2,4,5-трихлор-1,3-динитробензол (-)

ЗАКЛЮЧЕНИЕ

Таким образом, метод ДСМ позволяет в большинстве случаев сделать правильный прогноз мутагенности химических соединений. Сгенерированные методом ДСМ положительные гипотезы совпали с определенными экспертами структурными фрагментами – «тревожными сигналами» мутагенности. Выделены дополнительные «тревожные сигналы», определяющие мутагенность в тесте Эймса на штаммах TA100 и TA98, а также мутагенность во всех вариантах теста или хотя бы в одном из вариантов. Следует отметить, что мутагенная активность большинства соединений ряда замещенных нитробензолов обусловлена биоактивацией в ходе метаболических биотрансформаций, в частности реакцией восстановления нитрогруппы. Для таких соединений наличие мутагенности связано с легкостью осуществления метаболической трансформации, в которой участвует структурный фрагмент-гипотеза (например нитрогруппа), если эта трансформация приводит к образованию высоко реакционноспособных метаболитов, образующих аддукты с ДНК. Характеристика энергетических параметров реакций биотрансформации и реакционной способности метаболитов может быть осуществлена с использованием методов квантовой химии. Для правильного прогноза необходимо провести тщательный анализ информации по метаболизму

соединений под действием ферментных систем. Учет данных о механизме биотрансформаций, а также использование электронных параметров, характеризующих реакции биотрансформации и реакционную способность образующихся метаболитов в реакциях с ДНК, позволяет свести к минимуму ошибки прогноза.

СПИСОК ЛИТЕРАТУРЫ

1. OECD guideline for testing of chemicals. Bacterial Reverse Mutation Test. – URL: <http://www.oecd.org/chemicalsafety/risk-assessment/1948418.pdf> Accessed 10 April 2014.
2. Toxnet. Toxicology and Environmental Information. – URL: <http://toxnet.nlm.nih.gov/>
3. ISSSTY database on *Salmonella typhimurium* (Ames test) bacterial mutagenicity. – URL: <http://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?CCRIS>
4. D'yachkov P.N., Kharchevnikova N.V., Kuznetsov A.V., Dmitriev, A.V., Poroikov V.V. Quantum chemical simulation of cytochrome P450 catalyzed aromatic oxidation: Metabolism, toxicity, and biodegradation of benzene derivatives // *International Journal of Quantum Chemistry*. – 2007. – Vol. 107, № 13. – P. 2454-247.
5. Marques de Oliveira Iu., Bonatto D., Pêgas Henriques J. A. Nitroreductases: Enzymes with Environmental, Biotechnological and Clinical Importance //
6. Watanabe M., Nishino T., Koji T. Nitroreductases of Wild-type and Mutant "Classical" *Salmonella typhimurium* L33R Mutation Greatly Diminishes the binding of FMN to the Nitro-reductase // *J. Biol. Chem.* – 1998. – Vol. 273. – P. 23922-23928.
7. Nivinskas H, Koder R.L., Anusevicius Z., Sarlauskas J., Miller A.F., Cenas N. Quantitative Structure–activity relationships in two-electron reduction of nitroaromatic compounds by *Enterobacter cloacae* NAD(P)H nitroreductase // *Arch Biochem Biophys.* – 2001. – Vol. 385. – P. 170–178.
8. Haynes C.A., Koder R.L., Miller A.-F., Rodgers D. W. Structures of Nitroreductase in Three States. Effects of Inhibitor Binding and Reduction // *J. Biolog. Chem.* – 2002. – Vol. 277, № 13. – P. 11513–11520.
9. Yanto Y., Hall M., Bommaris A.S. Nitroreductase from *Salmonella typhimurium*: characterization and catalytic activity // *Org Biomol Chem.* – 2010. – Vol. 8, № 8. – P. 1826-1832.
10. Debnath A.K., Debnath G., Shusterman A.J., Hansch C. A QSAR investigation of the role of hydrophobicity in regulating mutagenicity in the Ames test.: I. Mutagenicity of aromatic and heteroaromatic nitro compounds in *Salmonella typhimurium* TA98 and TA100. // *Env. Mol. Mutag.* – 1992. – Vol.19. – P.53-70.
11. Mekenyan O., Dimitrov S., Serafimova R. et al. Identification of the Structural Requirements for Mutagenicity by Incorporating Molecular Flexibility and Metabolic Activation of Chemicals I: TA100 Model // *Chem. Res. Toxicol.* – 2004. – Vol. 17. – P. 753-766.
12. Serafimova R., Todorov M., Pavlov T., Mekenyan O. et al. Identification of the Structural Requirements for Mutagenicity, by Incorporating Molecular Flexibility and Metabolic Activation of Chemicals. II. General Ames Mutagenicity Model // *Chem. Res. Toxicol.* – 2007. – Vol. 20. – P. 662-676.
13. Kuz'min V.E., Muratov E.N., Artemenko A.G. et al. The effects of characteristics of substituents on toxicity of the nitroaromatics: HiT QSAR study // *J. Comp Aid Mol Des.* – 2008. – Vol. 22. – P. 747–759.
14. Enoch S.J., Cronin M.T.D. Development of new structural alerts suitable for chemical category formation for assigning covalent and non-covalent mechanisms relevant to DNA binding // *Mutat. Res.* – 2012. – Vol. 743. – P. 10– 19.
15. Enoch S.J., Cronin M.T.D. A review of the electrophilic reaction chemistry involved in covalent DNA binding // *Critical Reviews in Toxicology.* – 2010. – Vol. 40, № 8. – P. 728–748.
16. Финн В.К. Правдоподобные рассуждения в интеллектуальных системах типа ДСМ. // *Итоги науки и техники. Сер. Информатика.* – М.: ВИНТИ, 1991. - Т.15. - С. 54-101.
17. Blinova V.G., Dobrynin D.A., Finn V.K., Kuznetsov S.O., Pankratova E.S. Toxicology Analysis by Means of the JSM-method // *Biometrics.* – 2003. – Vol. 19. – P.1201-1207.

Материал поступил в редакцию 20.01.15.

Сведения об авторах

ХАРЧЕВНИКОВА Нина Вениаминовна – доктор биологических наук, старший научный сотрудник Научно-исследовательского института экологии человека и гигиены окружающей среды им. А.Н. Сысина Министерства здравоохранения Российской Федерации, Москва
e-mail: kharchevnikova_n@mail.ru

ЖУРКОВ Вячеслав Серафимович – доктор биологических наук, профессор, ведущий научный сотрудник Научно-исследовательского института экологии человека и гигиены окружающей среды им. А.Н. Сысина Министерства здравоохранения Российской Федерации
e-mail: vszhurkov@mail.ru

БЛИНОВА Валентина Георгиевна – кандидат технических наук, старший научный сотрудник Всероссийский институт научной и технической информации Российской академии наук
e-mail: blinova@viniti.ru

ДОБРЫНИН Дмитрий Анатольевич – кандидат технических наук, научный сотрудник, Всероссийский институт научной и технической информации Российской академии наук
e-mail: dobr@viniti.ru

Е.В.Громова, А.С.Денисов, А.В. Кочетков

Библиометрический анализ журналов, в которых публикуются преподаватели технических университетов

Проведен сравнительный анализ журналов, в которых опубликованы статьи сотрудников автомеханического факультета СГТУ имени Гагарина Ю.А. за период 2008-2012 гг., с использованием основных библиометрических индикаторов, а также рассмотрен механизм влияния этих параметров на публикационную активность.

Ключевые слова: публикационная активность, технические журналы, периодические издания, импакт-фактор, индикаторы, индекс цитирования, самоцитирование

Для оценки эффективности научной и инновационной деятельности организаций, научных коллективов, специалистов и ученых применяются различные индикаторы, в том числе показатели публикационной активности [1]: цитируемость, общее число публикаций организации, отраженных в БД «Российский индекс научного цитирования» (РИНЦ), количество научных статей, опубликованных за определенный период времени, индекс Хирша, импакт-фактор (ИФ) журналов, а также число публикаций: в зарубежных журналах, в российских журналах, в российских журналах из перечня ВАК, в российских переводных журналах и в журналах, входящих в Web of Science или Scopus, в журналах с импакт-фактором больше нуля; и кроме того число публикаций, процитированных хотя бы один раз, выполненных в сотрудничестве с другими организациями, с участием зарубежных авторов, количество авторов, количество авторов, имеющих публикации в журналах, входящих в Web of Science или Scopus и др.

Эти индикаторы включены в рейтинговую оценку эффективности работы профессорско-преподавательского состава и сотрудников Саратовского государственного технического университета имени Гагарина Ю.А. (СГТУ). Поэтому преподавателям и сотрудникам необходимо правильно и своевременно ориентироваться на этапе выбора издания для опубликования своих научных статей.

В качестве критериев оценки авторитетности научных журналов используют количественные (периодичность выпуска, тираж, количество публикуемых за год статей, удельный вес отклоненных рукописей, удельное отношение тиража к числу потенциальных подписчиков, средний объем публикаций и средняя численность авторского коллектива, средний срок публикации статей, распределение

тиража журнала по странам и т.д.) и качественные (методы анкетирования читателей или экспертных оценок научными организациями и учеными) показатели [2].

Известны исследования оценки научных журналов с использованием методов библиографического анализа баз данных (БД) по цитированию:

1) Web of Science – WoS (Thomson Reuters, США, Филадельфия) – поисковая платформа, объединяющая реферативные базы данных публикаций, охватывающая более 12 тыс. научных журналов по всем отраслям знания и 120 тыс. материалов конференций, более 4400 сайтов. К недостаткам сервиса Web of Science можно отнести то, что учитываются статьи только на английском языке.

2) Scopus (издательство Elsevier, Нидерланды, Амстердам) – международная реферативная база данных, индексирующая около 20 тыс. наименований журналов, в том числе научно-технических, более пяти тысяч издателей.

3) РИНЦ (eLibrary, Россия) – национальная информационно-аналитическая система, содержащая более 2 млн публикаций российских авторов и информацию о цитировании этих публикаций из более 3000 российских журналов [3]. Это уже достаточно развитой кластер российской информационной инфраструктуры науки и образования.

Доля российских журналов, отражаемых в международных базах данных, продолжает оставаться низкой (в частности, в Scopus входит около 300 наименований журналов), невелико и число индексируемых в них российских публикаций. Если рассматривать тематическую составляющую, то в основном это естественные и инженерные науки, медицина.

В отношении РИНЦ следует отметить, что система активно развивается и позволяет получать под-

робные показатели, использующиеся в различных отчетных формах. База данных некоммерческая, что делает ее доступной для проведения различного рода библиометрических исследований [3, 4].

Цель нашей работы – анализ публикаций преподавателей и сотрудников автомеханического факультета СГТУ имени Гагарина Ю.А. по выбору объектов исследования; сравнение основных библиометрических показателей научных журналов, в которых помещены эти публикации, и оценка механизма влияния этих показателей на публикационную активность авторов.

В качестве информационной базы использовались ресурсы, размещенные на платформе Научной электронной библиотеки eLibrary (РИНЦ – <http://elibrary.ru>), данные БД «Публикации сотрудников СГТУ» (Научно-техническая библиотека (НТБ) СГТУ – <http://lib.sstu.ru>), БД «Scopus» (<http://scopus.com>).

Результаты поиска по БД Scopus показали, что ни один из журналов транспортной тематики, в которых публиковались авторы СГТУ по данным 2008-2012 гг., в эту БД не входит. Поэтому для определения основного исследуемого массива были использованы БД «Публикации сотрудников СГТУ» (НТБ СГТУ) и РИНЦ (eLibrary).

Научные и отраслевые журналы были отобраны по количеству опубликованных работ (не менее трех) в периодических изданиях за пять лет (2008-2012 гг.). Следующее условие – включение журнала в список ВАК и его регистрация в БД РИНЦ. Журналы, не имеющие детальной статистики в БД РИНЦ (отсутствие номеров, неполный годовой комплект и т.д.), были исключены.

Распределение публикаций по журналам показывает, что у наших авторов сформировался круг изданий, которым они отдают предпочтение (рис. 1). Это определяется авторитетом научных школ и состава редакционных коллегий. В число приоритетных вхо-

дят: два вузовских издания «Вестник Саратовского государственного технического университета» (Вестник СГТУ) и «Известия Волгоградского государственного технического университета» (Известия ВолГТУ), а также журналы «Научное обозрение», «Ремонт, восстановление и модернизация», «Автотранспортное предприятие», «Мир транспорта и технологических машин». Сравнивать научные и отраслевые журналы разумно в пределах однородных категорий общей тематики. Поэтому журналы были разделены на две подгруппы: техническую (9 журналов) и мультидисциплинарную (4 журнала).

Как один из главных показателей измерителей динамики развития журнала используется импакт-фактор.

Согласно требованиям РИНЦ импакт-фактор регулярно пересчитывается в конце первого полугодия и со временем меняется. Для сравнения были выбраны оценки пятилетнего импакт-фактора, согласно показателям РИНЦ (рис. 2 и 3).

Согласно представленным данным, увеличение импакт-фактора четырех мультидисциплинарных журналов характерно для всех журналов, кроме «Вестника Саратовского государственного технического университета». Наибольший импакт-фактор имеют вузовские издания – «Вестник Оренбургского государственного университета» и «Известия Волгоградского государственного технического университета».

Анализ импакт-фактора девяти технических журналов показал, что наивысший показатель имеют журналы: «Автомобильная промышленность», «Трение и смазка в машинах и механизмах», «Ремонт, восстановление и модернизация». Тенденция к росту этого показателя видна у журналов «Автотранспортное предприятие», «Тяжелое машиностроение» (см. рис. 3).

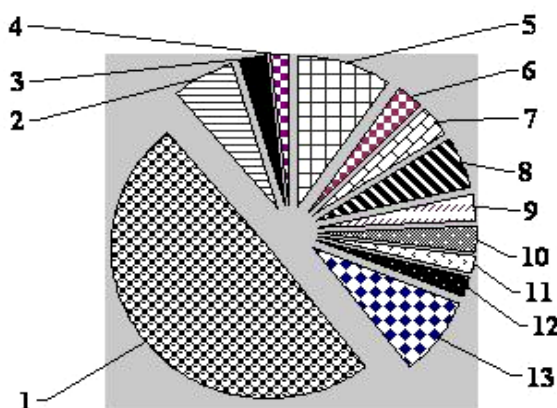


Рис. 1. Распределение публикаций преподавателей и сотрудников автомеханического факультета (АМФ) СГТУ в научных журналах (2008-2012 гг.):

- 1) «Вестник СГТУ» – 72 статьи; 2) «Известия ВолГТУ» – 9 статей; 3) «Строительные и дорожные машины» – 4 статьи; 4) «Вестник ОГУ» – 3 статьи; 5) «Автотранспортное предприятие» – 14 статей; 6) «Автомобильная промышленность» – 4 статьи; 7) «Мир транспорта и технологических машин» – 4 статьи; 8) «Ремонт, восстановление и модернизация» – 8 статей; 9) «Трение и смазка в машинах и механизмах» – 4 статьи; 10) «Грузовое и пассажирское автохозяйство» – 4 статьи; 11) «Грузовик» – 3 статьи; 12) «Тяжелое машиностроение» – 3 статьи; 13) «Научное обозрение» – 12 статей

Импакт-фактор

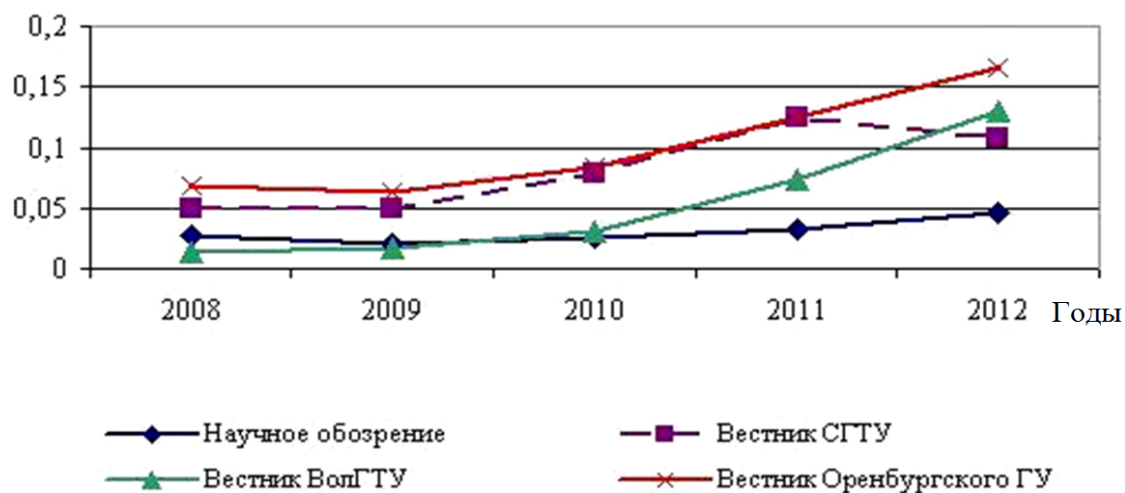


Рис. 2. Изменение пятилетнего импакт-фактора мультидисциплинарных журналов (данные на 10.04.2014 г.)

Импакт-фактор

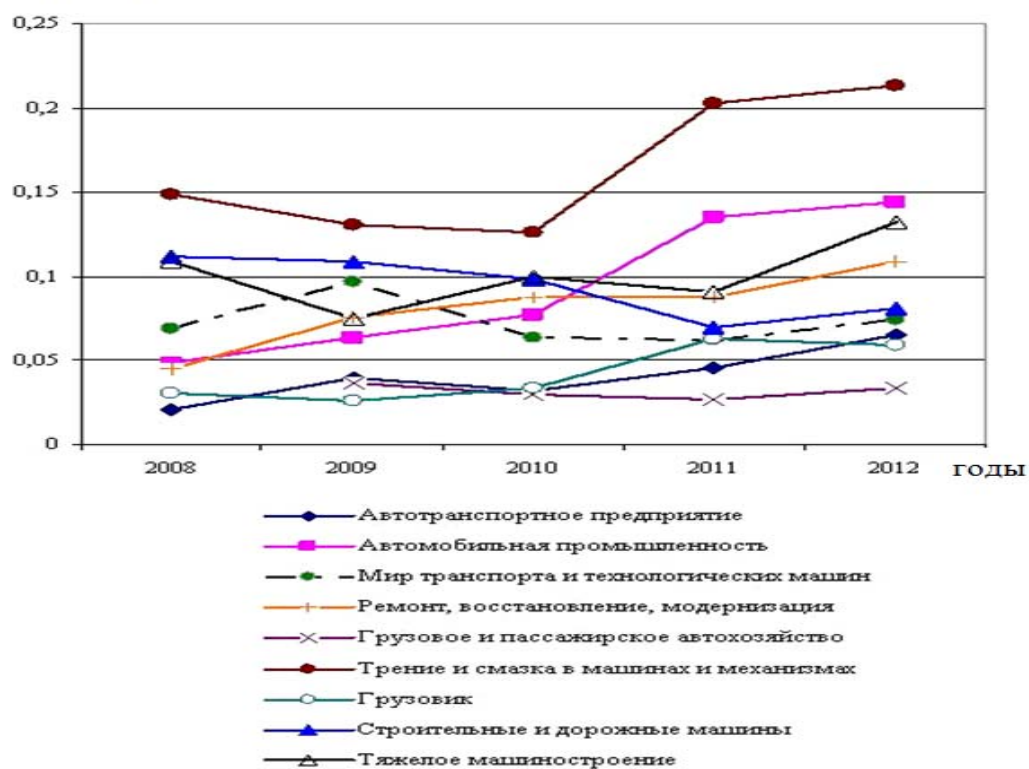


Рис. 3. Изменение пятилетнего импакт-фактора технических журналов (данные на 10.04.2014 г.)

Можно сделать вывод, что величины импакт-фактора журналов обеих групп различаются незначительно и лежат в близком диапазоне – 0,04 0,16. Именно этот показатель применяется в расчетах рейтинга профессорско-преподавательского состава СГТУ имени Гагарина Ю.А., поэтому авторы отдают предпочтение вузовским изданиям.

Величина импакт-фактора журнала напрямую зависит и от цитируемости его публикаций [5]. Установлено, что в группе мультидисциплинарных журналов (рис. 4) наиболее цитировались статьи «Научного обозрения» и «Вестника Оренбургского государственного университета».

Общее количество статей и суммарное число цитирований по данным РИНЦ (eLibrary.ru) для мультидисциплинарных журналов представлено в табл. 1. Общее количество статей и суммарное число цитирования по данным РИНЦ (eLibrary.ru)

для технических (автотранспорт) журналов представлено в табл. 2.

В группе журналов технического профиля наибольшее количество цитирований имели журналы «Автомобильная промышленность», «Тяжелое машиностроение», «Ремонт, восстановление и модернизация» и «Трение и смазка в машинах и механизмах» (рис. 5). Данные по цитированию журнала «Строительные и дорожные машины», представлены в РИНЦ не в полном объеме и в этих показателях не присутствуют. Между тем, этот журнал считается авторитетным и высоко оценивается при экспертизе качества и количества публикаций в диссертационных исследованиях. Наши данные подтверждаются позициями, занимаемыми этими журналами в рейтинге РИНЦ по предметной области «Машиностроение» и «Транспорт»: все вышеуказанные издания входят в первую половину из восьмидесяти представленных в этом рейтинге журналов.

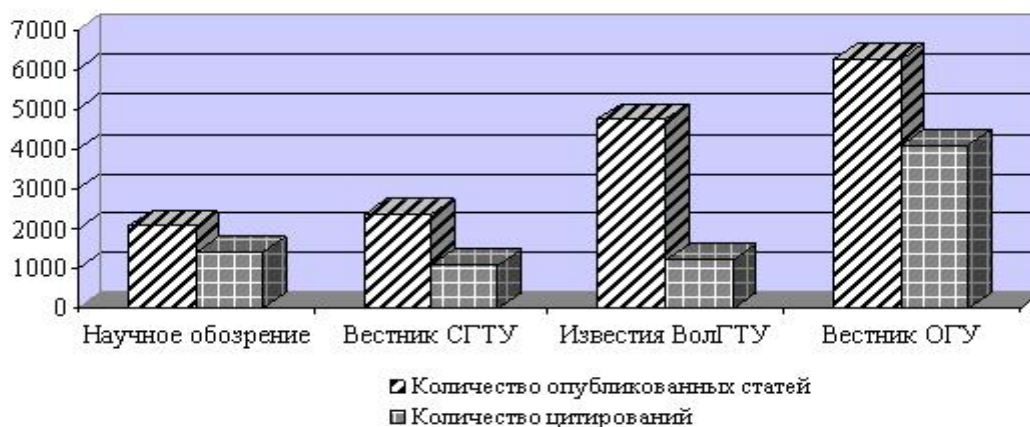


Рис. 4. Соотношение количества опубликованных статей и цитирований для мультидисциплинарных журналов

Таблица 1

Общее количество статей и суммарное число цитирования по данным РИНЦ (eLibrary.ru) для мультидисциплинарных журналов*

Показатели	Названия журналов			
	«Научное обозрение»	«Вестник Саратовского государственного технического университета»	«Известия Волгоградского технического университета»	«Вестник Оренбургского государственного университета»
Общее число статей, размещенных в РИНЦ (по данным РИНЦ)	2055	2346	4766	6284
Суммарное число цитирований (по данным РИНЦ)	1392	1097	1213	4100

*Данные на 10.04.2014 г. В конце 2014 г. показатели в РИНЦ были обновлены.

Общее количество статей и суммарное число цитирований по данным РИНЦ (eLibrary.ru) для технических (автотранспорт) журналов*

Показатели	Названия журналов							
	«Автотранспортное предприятие»	«Автомобильная промышленность»	«Мир транспорта и технологий машин»	«Ремонт, восстановление, модернизация»	«Грузовое и пассажирское хозяйство»	«Трение и смазка в машинах и механизмах»	«Грузовик»	«Тяжелое машиностроение»
Общее число статей, размещенных в РИНЦ (по данным РИНЦ)	1621	1729	290	1569	998	740	1671	1960
Суммарное число цитирований (по данным РИНЦ)	406	1071	104	924	98	504	251	1201

* Данные на 10.04.2014 г. В конце 2014 г. показатели в РИНЦ были обновлены.

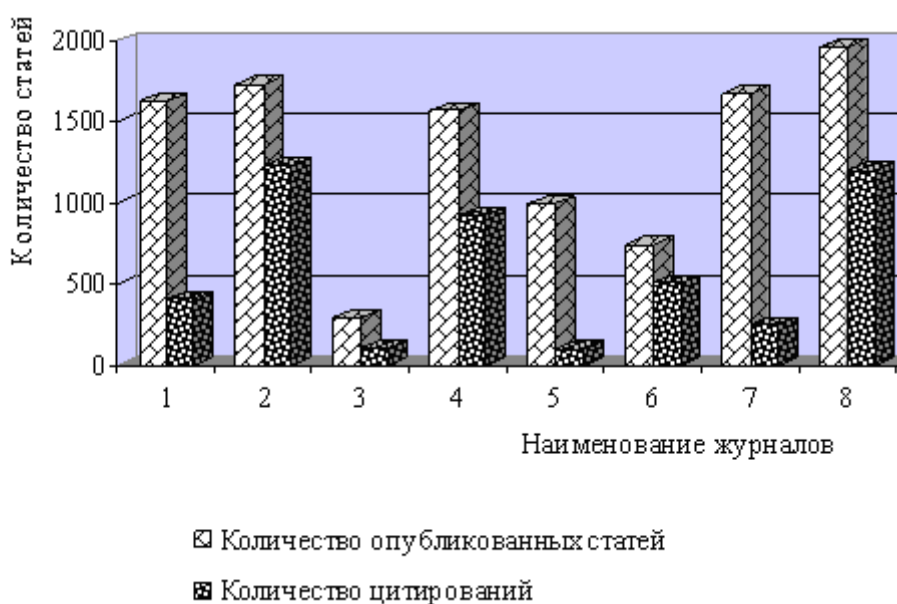


Рис. 5. Соотношение количества опубликованных статей и цитирований в группе технических журналов:

- 1) «Автотранспортное предприятие»; 2) «Автомобильная промышленность»; 3) «Мир транспорта и технологий машин»; 4) «Ремонт, восстановление, модернизация»; 5) «Грузовое и пассажирское автохозяйство»; 6) «Трение и смазка в машинах и механизмах»; 7) «Грузовик»; 8) «Тяжелое машиностроение»

Рост ряда библиометрических показателей следует рассматривать со снижением научной и имиджевой значимости журнала – это коэффициент самоцитирования и доля публикаций, не имеющих ни одной ссылки. Считается, что чем меньше коэффициент самоцитирования, тем весомее ссылки на журнал.

Для группы мультидисциплинарных журналов установлено, что наибольший показатель самоцитирования (от 30 до 55 %) характерен для «Вестника Саратов-

ского государственного технического университета», «Известий Волгоградского технического университета» и «Вестника Оренбургского государственного университета». Исключением является «Научное обозрение» (показатель самоцитирования не более 20%).

Журналы технического профиля имеют показатель самоцитируемости не более 40 %. Они распределяются следующим образом: «Автомобильная промышленность» – 40 %, «Трение и смазка в машинах и

механизмах» – 39 %, «Грузовик» – 35 %; «Ремонт, восстановление, модернизация» – около 30 %; «Автотранспортное предприятие» – 26 %; «Тяжелое машиностроение» – 25 %. «Мир транспорта и технологических машин» – 20 %; «Грузовое и пассажирское автохозяйство» – 13,8 %.

Установлено, что в открытом доступе находятся только вузовские журналы, входящие в мультидисциплинарную группу. Для журналов технического профиля полные тексты оказались доступны только для подписчиков на сайтах издательств и в НЭБ eLibrary. Исключение составляют журналы «Автомобильная промышленность», «Грузовик», «Трение и смазка в машинах и механизмах». Издательство предоставляет свободный доступ к архивам своих журналов за предыдущие годы со значительным запаздыванием. Это ограничивает возможности читателей быть в курсе новых разработок и обратной связи с автором в виде, в том числе, цитирования.

Важен еще один библиометрический показатель – среднее количество источников в списках пристатейной литературы. В авторитетных зарубежных журналах среднее число пристатейной литературы составляет около 40-50 источников. В среднем на 34 источника ссылаются в своих работах российские авторы, публикующиеся за рубежом. Авторы, публикующиеся в российских журналах, в пристатейных ссылках используют максимум 10-12 источников [6].

Анализ списков пристатейной литературы в рассматриваемых журналах показал, что число цитируемых источников в среднем составляет четыре-шесть. Основное количество цитируемой литературы – это статьи, книги, патенты, сборники научных статей, материалы конференций, учебные издания.

Ни один из рассматриваемых нами журналов двух групп в настоящее время не включен в международные БД по цитированию, в частности, в базу Scopus.

Чтобы российские журналы стали полноправными участниками этих ресурсов недостаточно иметь индекс ISSN и публиковать рефераты на английском языке. Существует строгая многоступенчатая экспертиза по включению журналов в международные базы данных. В настоящее время ситуация стала меняться.

Хотелось бы надеяться, что многие издательства уже ориентированы на работу в этом направлении, что позволит поднять авторитет наших журналов на качественно новую ступень.

ВЫВОД

Основное количество журнальных публикаций авторов автомеханического факультета Саратовского государственного технического университета имени Гагарина Ю.А. и Пермского национального исследовательского политехнического университета приходится на вузовские журналы (в частности, в «Вестнике Саратовского государственного технического университета» опубликовано более 50% статей). Анализ показал, что такие журналы имеют невысокий импакт-фактор, узкую цитируемость, не имеют международного признания.

Статьи в подобных журналах позволяют авторам повышать количественные показатели публикацион-

ной активности (рост публикаций), получать баллы в ежегодном университетском рейтинге, однако не увеличивают показатели цитируемости. Поэтому авторы должны более требовательно подходить к выбору журналов для публикации своих работ.

Авторам – сотрудникам автомеханического факультета СГТУ необходимо расширить круг изданий для своих статей, обратив внимание на журналы, имеющие более высокий импакт-фактор, а также, рассмотреть возможность публикации в журналах, входящих в международные базы по цитированию.

СПИСОК ЛИТЕРАТУРЫ

1. Мохначева Ю. В., Харибына Т. Н. Методика определения значимости научных публикаций // Библиосфера. – 2008. – № 3. – С.23-33.
2. Шарабчив Ю. Т. Оценка значимости научных журналов и ученых по цитируемости // Медицинские новости. – 2010. – № 8. – С. 6-9.
3. Полникова Е. М., Шабанова С. М. Научная электронная библиотека eLIBRARY.RU: Руководство пользователя, ООО «РУНЭБ»; СПб. гос. ун-т. – М. : РУНЭБ, 2010. – 48 с.
4. БД РИНЦ – российский индекс научного цитирования // Научная электронная библиотека eLibrary. – URL: <http://elibrary.ru> (дата обращения к ресурсу: 24.03.2014).
5. Елин А. Л., Елин А. Л., Шапошников Ю. Ю. Заметки к вопросу об эффективности использования различных наукометрических показателей и критериев эффективности научных исследований // Научная периодика: проблемы и решения. – 2013. – № 03. – С. 4-12.
6. Еременко Г. О. Российский индекс научного цитирования и информационно-аналитическая система. Систем. требования: Ms. Office: Power Point. – URL: http://elibrary.ru/projects/science_index/conf/2012/presentations/eremenko1.pdf. (дата обращения 30. 01. 2013).

Материал поступил в редакцию 05.01.15.

Сведения об авторах

ГРОМОВА Елена Владимировна – зав. отделом Научно-технической библиотеки, Саратовский государственный технический университет имени Гагарина Ю.А.
e-mail: yagromovaev@mail.ru

ДЕНИСОВ Александр Сергеевич – доктор технических наук, профессор Саратовского государственного технического университета имени Гагарина Ю.А., эксперт ВАК Минобрнауки
e-mail: denisov0307@yandex.ru

КОЧЕТКОВ Андрей Викторович – доктор технических наук, профессор Пермского национального исследовательского политехнического университета
e-mail: soni.81@mail.ru

АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 81'32 : 811.112.28

Т.А. Архангельский, О.А. Созинова

Мультимедийный корпус языка идиш*

Представлен созданный авторами мультимедийный корпус языка идиш. Первая версия корпуса содержит 10 часов аудио- и видеоматериалов, синхронизированных с расшифровками. Для корпуса были созданы поисковая платформа и веб-интерфейс, использующие базу данных типа NoSQL, веб-фреймворк Django и ряд модулей на языке JavaScript. Веб-интерфейс позволяет выполнять лексико-грамматические запросы и просматривать результаты в виде расшифровки, подсветка которой синхронизирована с проигрываемым мультимедийным материалом. Описаны отличия мультимедийного корпуса идиша от аналогичных мультимедийных корпусов и преимущества созданной поисковой платформы.

Ключевые слова: идиш, мультимедийный корпус, документация языков

ВВЕДЕНИЕ

Идиш является крайне интересным объектом исследования для лингвистов по целому ряду причин. Лексика идиша включает в себя несколько пластов: германский, славянский и семитский; идиш имеет большое количество диалектов, которые характеризуются значительными различиями в фонетике и грамматике; морфология и синтаксис идиша из-за долгих языковых контактов приобрели ряд славянских черт. Однако именно лингвистических работ, посвящённых исследованию идиша, крайне мало. Во многом это связано с отсутствием материалов и инструментов для исследования. Единственный корпус языка идиш (<http://web-corpora.net/YNC/search/>) появился лишь несколько лет назад и включает только письменные тексты [1]; для анализа современной устной речи исследователям необходимо искать информантов (что особенно сложно, если требуется получить данные нескольких диалектов).

Целью нашей работы было создание мультимедийного корпуса языка идиш, т.е. собрания расшифрованных видео- и аудиозаписей с грамматической разметкой и возможностью поиска посредством онлайн-интерфейса. Мультимедийные корпуса языков – сравнительно недавнее явление, и немногие языки обладают в настоящий момент таким ресурсом. Это связано с тем, что составление подобных корпусов требует больших затрат времени и сил, чем создание корпуса письменного языка, из-за меньшей доступности видео- и аудиозаписей, трудоёмкости расшиф-

ровки и более сложного устройства такого корпуса с технической точки зрения. Мы надеемся, что создание такого ресурса позволит увеличить количество исследований, использующих идишский материал.

Описываемый проект включал в себя сбор и расшифровку текстов, разработку принципов функционирования онлайн-интерфейса корпуса и создание поисковой платформы корпуса. В настоящий момент корпус содержит около 10 часов расшифрованных мультимедийных материалов, завершена разработка первой версии поисковой системы и онлайн-интерфейса, который будет размещён в Интернете по адресу: <http://web-corpora.net/YiddishMultimediaCorpus/search/>

ОБЗОР МУЛЬТИМЕДИЙНЫХ КОРПУСОВ

Первым этапом работы, предшествующим разработке собственной поисковой платформы мультимедийного корпуса, стало изучение существующих онлайн-интерфейсов мультимедийных корпусов. Нас интересовали технологии, которые использовались при создании корпусов, а именно возможности поиска, реализация проигрывания видео- и аудиофайлов и синхронизация мультимедийного содержимого с текстом. Всего было рассмотрено 10 корпусов, находящихся в свободном доступе в Интернете.

Из рассмотренных корпусов наиболее мощным поиском обладает мультимедийный подкорпус Национального корпуса русского языка [2] (<http://www.ruscorpora.ru/search-murco.html>). Разметка корпуса позволяет задать не только грамматические характеристики, но и ряд других параметров, например, количество говорящих в клипе, типы речевых действий, ориентацию ладони, направление движения и многое другое. В корпусе использован язык

* Данное научное исследование (14-05-0074) выполнено при поддержке Программы «Научный фонд НИУ ВШЭ» в 2014 г.

программирования JavaScript, видео воспроизводится с помощью плеера Yandex Player 13.101-123 в формате FLV, есть возможность скачать видео в формате MP4. Текст выдачи можно расширить, чтобы посмотреть на контекст. Синхронизации текста выдачи с видеофайлом нет.

Одним из самых первых и известных мультимедийных корпусов является корпус детской речи CHILDES, являющийся частью системы TalkBank [3]. Этот корпус содержит записи детской речи, самые ранние из которых относятся к 1960-м годам, на нескольких языках. В онлайн-интерфейсе корпуса CHILDES Transcript Browser (<http://childes.psy.cmu.edu/browser/index.php>) для поиска используется специально разработанный для этого проекта язык запросов CLAN. На сайте находятся директории с файлами различного типа: аудио-, видео- и текстовые файлы. Видеоматериалы воспроизводятся при помощи плагина проигрывателя QuickTime, который должен быть предварительно установлен у пользователя. В тексте выдачи все строки пронумерованы для синхронизации видео с текстом.

Ещё один корпус, использующий плагин плеера QuickTime для воспроизведения аудиофайлов, – Scottish Corpus of Texts & Speech [4] (<http://www.scottishcorpus.ac.uk/>). Кроме простого текстового поиска в этом корпусе предусмотрен поиск по метаинформации: имеется возможность задать автора, тип текста и некоторые другие параметры с помощью чекбоксов. Аудиофайл открывается в новом всплывающем окне, текст синхронизирован с аудиодорожкой. И текст, и аудиофайл можно скачать. Корпус написан на PHP и JavaScript.

Устные корпуса испанского языка Spanish Learner Language Oral Corpora (<http://www.splloc.soton.ac.uk/search.php>) и французского языка French Learner Language Oral Corpora (<http://www.flloc.soton.ac.uk/search.php>) построены по технологии корпуса CHILDES. В корпусе предусмотрен простой текстовый поиск с возможностью задать несколько параметров. Аудио открывается в том же окне. Корпуса написаны с использованием JavaScript.

Корпус жестов National Center for Sign Language and Gesture Resources (NCSLGR) corpus [5] (<http://secrets.rutgers.edu/dai/queryPages/querySelection.php>) обладает только простым поиском с использованием чекбоксов по разным свойствам жестов. В качестве результата поиска выдаётся таблица с названиями жестов и именами людей. В пересечении строк и столбцов находится ссылка на видео с жестом, а также на видео с контекстом. Поисковая страница написана на PHP, также использован JavaScript. Видео проигрывается при помощи встроенного плеера JW Player 5.1.897.

Мультимедийный корпусный ресурс с одним из самых красивых интерфейсов, Corpus Project in Colloquial Japanese Sign Language [6] (<http://research.nii.ac.jp/jsl-corpus/en/>), почти полностью написан на Flash. По сути, корпус представляет собой базу данных видео с жестами без возможности поиска. Видео появляется во всплывающем окне, во время воспроизведения аннотация к жестам появляется внутри самого видео.

В Диалектологическом корпусе Бассейна реки Устья (Архангельская область, Устьянский район)

(<http://www.slavist.de/Pushkino/login.php>) есть три поля для поиска: простой поиск, лексико-грамматический поиск и поиск с помощью CQP-запросов. В выдаче присутствуют аудиофайлы и текст. Результаты поиска можно экспортировать в форматах CSV и XML, аудиофайлы можно скачать в формате WAV. В корпусе есть возможность получить расширенный контекст для результатов запроса, однако аудио при этом не изменяется. Синхронизации аудио с текстом нет. Корпус полностью написан на PHP.

Педагогический корпус BACKBONE Pedagogic corpora for content & language integrated learning [7] (<http://webapps.ael.uni-tuebingen.de/backbone-search/faces/search.jsp>), созданный в Университете Тюбингена, является базой данных, которая содержит видеointервью с носителями шести языков. Видео- и аудиофайлы нельзя проиграть на сайте, их можно только скачать (формат WVX). Поиск осуществляется по категориям в базе данных, а также по коллокациям. Корпус написан с использованием JavaScript.

Ещё один проект Университета Тюбингена – English Language Interview Corpus as a Second-Language Application (http://www.uni-tuebingen.de/elisa/html/elisa_index.html) содержит интервью с носителями английского языка. По реализации этот корпус очень похож на BACKBONE. В нём также возможно только скачивание видео (формат SMIL) и просмотр текста. Есть возможность загрузить результаты в формате XML, а также просмотреть частотные списки слов.

Как видно из обзора, большинство корпусов даёт возможность простого текстового поиска, а также поиска по метаинформации и дополнительной аннотации, тип которой зависит от предназначения корпуса (например, поиск по жестам). Большинство корпусов не имеет синхронизации текста и проигрываемого аудио- или видеофрагмента, а некоторые из них позволяют только скачивать мультимедийное содержимое в виде отдельных файлов, не предоставляя возможности проигрывания его на странице.

Проанализировав достоинства и недостатки существующих платформ для мультимедийных корпусов, мы решили руководствоваться следующими принципами.

- В качестве языка клиентских скриптов будет использоваться JavaScript; для удобства обработки запросов и результатов в онлайн-интерфейсе при обмене информацией между модулями будет использоваться формат JSON, поддерживаемый языком JavaScript.
- В интерфейсе не будут использоваться Flash и другие технологии, требующие установки дополнительных плагинов к браузеру, существующие не для всех систем.
- Корпус будет содержать встроенный проигрыватель. Это продиктовано как соображениями удобства (пользователь может сразу просмотреть видео, относящееся к найденному фрагменту расшифровки), так и соображениями авторского права: мы не имеем возможности предоставить пользователям для скачивания полные версии некоторых видеофайлов.
- Текст выдачи будет двусторонне синхронизирован с видео: при проигрывании в тексте будет под-

свечиваться расшифровка проигрываемого фрагмента, а при нажатии на какой-либо фрагмент будет запускаться соответствующий ему фрагмент видео.

ХАРАКТЕРИСТИКИ КОРПУСА

Перед началом разработки корпусной платформы мы сформулировали требования, которым корпус должен удовлетворять с точки зрения поисковых возможностей, принимая в расчёт предполагаемую целевую аудиторию корпуса и задачи, для решения которых он будет использоваться. Поскольку все расшифровки аудио- и видеоматериалов, составляющих мультимедийный корпус, также будут содержаться и в Корпусе языка идиш («основном» корпусе), мы посчитали разумным выделить те группы задач, которые невозможно решить без аудио- и видеосопровождения, и при разработке мультимедийного корпуса ориентироваться в первую очередь на эти задачи.

Предполагаемыми пользователями корпуса являются исследователи, которых интересует просодия, фонетика, жесты и социолингвистика идиша. Для проведения корпусных исследований в этих областях необходимо иметь возможность параллельного просмотра видеоматериала и сопровождающей его расшифровки. При этом в корпусе должен быть реализован поиск по расшифровкам. Для проведения типологических и грамматических исследований хорошо приспособлен основной корпус, оснащённый мощным поиском. В большинстве описанных выше случаев, однако, хватает простого поиска по тексту или по лемме. Метаинформация же в мультимедийном корпусе должна быть более детальной: например, без информации о диалекте говорящего будет невозможно осуществлять социолингвистические и диалектологические исследования (разница в произношении отдельных слов носителями разных диалектов), а без детального описания жанра и типа текста станет сложным исследование жестов.

Исходя из этих соображений, мы решили реализовать в первой версии корпуса следующий поисковый функционал:

1. Поиск по словоформе (с использованием регулярных выражений).
2. Поиск по лемме (с использованием регулярных выражений).
3. Поиск по грамматике (с использованием логических функций).
4. Поиск по метаинформации.

При одновременном задании нескольких параметров они должны объединяться в один запрос с помощью конъюнкции. Поиск конструкций, состоящих из нескольких слов, реализованный в основном корпусе, в мультимедийном корпусе реализован не будет, так как он необходим в первую очередь для грамматических исследований.

Как и в основном корпусе, мы не имеем возможности предоставлять пользователям доступ к полным версиям аудио- и видеоматериалов. Одной из причин являются авторские права: авторы видео дают разрешение использовать видео в корпусе, но не размещать его целиком. Другая причина – ограниченные серверные мощности и пропускная способность ка-

нала, которых может не хватить, если поисковая выдача будет включать полные версии видео. Тем не менее, при исследовании устной речи обычно требуется доступ к более широкому контексту каждого результата поиска, чем при работе с письменными текстами. Мы решили эту проблему разбиением видеофайлов на пересекающиеся фрагменты длиной 3 минуты тремя разными способами: с отступом 0, 1 и 2 минуты. В итоге каждый кадр исходного видео содержится в трёх разных фрагментах и как минимум в одном из них он имеет отступ не менее одной минуты от начала и конца фрагмента. При индексации расшифровок каждому сегменту расшифровки присписывается адрес фрагмента видео, оптимального с точки зрения широты контекста. Если полная версия видео доступна на сайте его авторов или на каком-либо видеохостинге, результаты поиска сопровождаются ссылкой на полную версию.

ТЕКСТЫ

В первую версию корпуса включено около 10 часов мультимедийного материала, основную часть которого составляют экспедиционные записи (юго-восточный диалект) и лекции носителей идиша (северо-восточный диалект). Поскольку было решено синхронизировать текст с видео, при расшифровке было необходимо указывать, к какому фрагменту видео относится расшифровываемый фрагмент. Для этого была использована система ELAN, а итоговые файлы расшифровок хранятся в XML-формате eaf. Разметка фрагментов для привязки к видео преследовала исключительно практическую цель обеспечения синхронизации в корпусе и поэтому в общем случае не отражает какого-либо лингвистически осмысленного деления текста (например, на предложения или элементарные дискурсивные единицы), хотя в большинстве случаев границей фрагментов является пауза. Каждый фрагмент имеет длину несколько секунд.

Орфография

При выборе правил расшифровки мы руководствовались, во-первых, спецификой материала, а во-вторых, поисковыми задачами корпуса.

Специфика идишской устной речи по сравнению с письменной состоит, во-первых, в большей вариативности и частом проявлении диалектных черт в фонетике и в лексике, а во-вторых, во взаимодействии идиша с другими языками (например, русским). Последнее выражается в обилии спонтанных заимствований и переключений кодов и вызвано тем, что носители, фигурирующие в видеоматериалах, владеют и другими языками как родными, а для некоторых участников диалогов – например, интервьюеров или слушателей публичных лекций – идиш вообще не является родным, что побуждает носителей переключаться на общий для говорящих и слушающих код.

Задачи корпуса как поискового инструмента, с другой стороны, требуют сведения вариативности к минимуму. При поиске какого-либо слова оно должно находиться во всех диалектных и фонетических вариантах, поскольку одна из основных задач корпуса – предоставить пользователям возможность сравнить

произнесение одного и того же слова носителями разных диалектов. Этого можно добиться, если каждая словоформа будет записана в корпусе в соответствии с правилами литературного языка. Ещё одним обстоятельством, требующим наличия стандартизированной записи, является использование морфологического парсера. Парсер, разработанный в рамках проекта «Корпус языка идиш» Программы фундаментальных исследований Президиума РАН «Корпусная лингвистика», работает с литературным языком и не способен распознавать диалектные варианты словоформ. Одним из решений было бы сопровождение точной расшифровки «нормализованным» вариантом, который был бы представлен в виде дополнительного слоя информации, выровненного с первым. Такое решение предлагалось, например, в [8] для русского диалектного корпуса. Этот подход позволяет сохранить больше информации, но является намного более ресурсоёмким и не имеет серьёзных преимуществ при наличии аудио- и видеосопровождения расшифровок. Поэтому мы, вслед за создателями текущей версии Диалектного подкорпуса НКРЯ [9] и аудиокорпуса Устьянского говора [10], решили придерживаться орфографического подхода: при расшифровке во всех возможных случаях словоформы записываются в стандартизированном виде, соответствующем литературной норме.

В случае с идишем, однако, такое решение требует дополнительных уточнений. Дело в том, что на практике для записи идишских текстов существуют десятки различных орфографических систем (см. [11] об истории идишского правописания), основанных на еврейском письме. Стандартный вариант орфографии был предложен институтом YIVO в 1937 г., однако он до сих пор не соблюдается многими издательствами, а кроме того, не даёт исчерпывающего описания всей орфографии, оставляя в некоторых случаях выбор за пишущим. Также институтом YIVO была разработана стандартная система транслитерации латиницей.

При расшифровке мы остановились на использовании стандартной латинской транслитерации, а не одной из орфографий, использующих еврейскую письменность. Причин такому решению несколько. Во-первых, использование транслитерации не только не приводит к потере информации, но и, наоборот, добавляет новую информацию: если в стандартной орфографии гебраизмы – заимствования из иврита и арамейского – пишутся по правилам языка-оригинала, без огласовок, то транслитерация позволяет их прочитать и тем, кто не знаком с этим пластом идишской лексики. Во-вторых, применение стандартной орфографии может быть затруднено при записи некоторых заимствований, которыми изобилуют тексты. В-третьих, такой подход позволяет передавать переключения кодов с помощью переключения систем письма (например, русский текст может быть записан кириллицей), что вызвало бы проблемы при использовании оригинальной орфографии из-за разницы в направлении письма.

Основываясь на описанной концепции, мы составили инструкцию для разметчиков, в которой были более подробно рассмотрены отдельные пункты. Ни-

же приведены основные решения, которые было необходимо разработать для сложных случаев:

1. В расшифровках не приводятся к литературной форме слова, имеющие существенные грамматические отличия от литературного идиша, например, использующие другую морфему для выражения какого-либо значения.

2. Если в диалекте одной литературной лексеме соответствует два варианта с разным значением, это различие отмечается. Например, слову *shul* ‘школа; синагога’ в некоторых диалектах со стандартным фонетическим переходом *u* → *i* соответствует пара слов *shil* (с переходом) и *shul* (без перехода), одно из которых имеет значение «школа», а другое — «синагога». В таком случае различие *shul/shil* должно сохраняться и, соответственно, форма *shil* не должна нормализовываться.

3. Переключения кодов планируется записывать в орфографии языка, на который переключается говорящий (в частности, переключения на русский, довольно частые в нашем корпусе, записываются кириллицей). Однако заимствования зачастую бывает сложно или невозможно отличить от переключения кодов, из-за чего некоторые исследователи даже предлагают отказаться от этого различия [12]. Следовательно, для достижения согласованности расшифровщиков необходимо разработать чёткие правила, позволяющие выделить переключения кодов. Такими при расшифровке считаются только фрагменты длиной более одного слова, которые невозможно интерпретировать как заимствования по грамматическим или фонологическим соображениям.

АРХИТЕКТУРА ПОИСКОВОЙ ПЛАТФОРМЫ

Поисковая платформа разрабатывалась в соответствии с требованиями, сформулированными в п. 2. Архитектура платформы включает три основные части: пользовательский интерфейс, выполненный на HTML/JavaScript, поисковый механизм, включающий в себя базу данных и модуль на языке Python, через который можно задавать поисковые запросы и получать результаты, и веб-сервер, написанный также на языке Python с использованием веб-фреймворка Django. Последний отвечает, среди прочего, за связь интерфейса с поисковым механизмом и авторизацию пользователей. Данные между этими тремя модулями передаются через json-объекты.

Веб-сервер

В функции веб-сервера входит, во-первых, выдача веб-страниц по запросу, а во-вторых, организация взаимодействия между интерфейсом и поисковой системой. Сервер корпуса, как и поисковый механизм, написан на языке Python с использованием одного из самых распространённых веб-фреймворков – Django. Программа, написанная с использованием этого веб-фреймворка, может либо самостоятельно обслуживать запросы, либо работать через посредство другого сервера, установленного в системе. Первой из этих возможностей мы пользовались во время тестирования и отладки интерфейса корпуса; версия,

доступная для пользователей, получает запросы и отправляет ответы через веб-сервер Apache.

Дополнительное звено между интерфейсом и поисковым механизмом позволило реализовать несколько функций, невозможных при использовании более простой структуры. Веб-сервер отвечает за следующие функции:

1. Преобразование поискового запроса пользователя в json-запрос к базе данных. Пользователь вводит информацию о словоформе, которую он хочет найти, в несколько текстовых полей. В настоящий момент это поля «Лемма», «Словоформа» и «Грамматика». Если первые два поля содержат обычные строки, то поле «Грамматика» может содержать сложные выражения, содержащие логические функции: конъюнкцию, дизъюнкцию и отрицание (например, «найти все глаголы в форме причастий или прилагательные»). Строка запроса в этом случае должна быть преобразована в json-объект древовидной структуры (см. п. 5.2), что удобнее делать на сервере, чем на компьютере клиента.

2. Получение ответа от поискового сервера и отправка его в интерфейс по частям. Поскольку по запросу пользователя может быть найдено очень много результатов, мы решили придерживаться стандартной схемы выдачи результатов, согласно которой они делятся на страницы (по умолчанию 20 результатов на странице). Если бы все результаты загружались сразу, это привело бы к большой нагрузке на сервер и чрезмерно долгому ожиданию ответа пользователем. Сервер, получив ответ от поискового механизма, отправляет часть полученных результатов и информацию об общем количестве результатов клиенту, после чего пользователь может запросить другую часть результатов, нажав на ссылку с номером страницы.

3. Поддержка нескольких языков интерфейса. Было решено выполнить интерфейс корпуса на двух языках – русском и английском (с перспективой добавить в дальнейшем интерфейс на идише). Наиболее удобный и масштабируемый вариант решения этой задачи состоит в том, что в HTML-коде страниц вместо надписей помещаются специальные шаблоны, на место которых сервер вставляет текст из соответствующего языкового файла.

4. Панель администрирования. В платформу было решено добавить несколько функций, доступных администраторам, таких как просмотр истории поисковых запросов пользователей и загрузка новых текстов. Фреймворк Django хорошо подходит для этой задачи, поскольку имеет встроенные средства для регистрации пользователей и создания панели администратора.

Поисковый механизм

В большинстве существующих корпусных поисковых систем используются либо реляционные базы данных (например, в Corpus of Contemporary American English и Corpus del Español [13] или в Восточно-армянском национальном корпусе [14] и других корпусах, использующих ту же систему, в том числе в основном идишском корпусе), либо специальный веб-сервер с индексными файлами, не хранящимися в

базе данных (например, в системе Corpus Workbench [15] и в Национальном корпусе русского языка [16]). В нашей поисковой системе мы решили пойти по третьему пути и использовать нереляционные (NoSQL) базы данных. Такой подход уже применялся на практике (например, CouchDB в комбинации с Apache Lucene в [17] и MongoDB в [18]), но ещё не получил большого распространения в создании корпусов из-за новизны этой технологии.

В качестве базы данных была выбрана MongoDB. В отличие от реляционных баз данных, в MongoDB единицей хранения является BSON-документ (аналог строки в реляционных базах); документы объединяются в коллекции (аналог таблицы). В отличие от реляционных баз, задающих жёсткие рамки в виде фиксированного числа и типа столбцов в каждой таблице, объекты в MongoDB могут иметь любое количество свойств и иметь сложную внутреннюю структуру. Поисковый запрос к MongoDB представляет собой json-объект специального вида. Несмотря на намного более свободную организацию хранения, нереляционные базы данных позволяют создавать индексы по любым полям объектов и их сочетаниям и использовать их для быстрого поиска.

В текущем варианте поисковой системы база данных содержит коллекции текстов, предложений и словоформ. Каждый документ в коллекции текстов содержит метаданные для одного из текстов корпуса. Документы коллекции слов содержат информацию обо всех уникальных словоформах, имеющихся в корпусе, вместе с их грамматическими разборами и всей остальной информацией, доступной для поиска. Эта коллекция играет роль обратного индекса: каждый документ содержит список идентификаторов предложений, в которых содержится соответствующая словоформа. Документы коллекции предложений описывают предложения (или другие элементарные фрагменты текста), представляющие собой массивы словоформ и содержащие дополнительную информацию, в частности, идентификаторы соседних предложений и идентификатор документа. Коллекция предложений, в отличие от двух других коллекций, предназначена в первую очередь для выдачи результатов, а не для поиска, поэтому словоформы в ней могут содержать дополнительные поля, содержимое которых должно быть показано в выдаче, но не должно быть доступно для поиска.

Ниже приводится начальный фрагмент одного из предложений коллекции:

```
{ "video_source": "video1.mp4",  
  "doc": "interview_1.eaf",  
  "start_offset": "1069.49",  
  "words": [  
    { "wf": "zaynen",  
      "type": "word",  
      "ana": [  
        { "lex": "zayn",  
          "transl_en": "be",  
          "gr": { "tense": "pres",  
                  "pos": "V",  
                  "number": "pl" } }  
      ]  
    }  
  ]  
}
```

```
{"wf": "linkse",  
  "type": "word"},  
{"wf": ",",  
  "type": "punc"},  
...
```

В примере выше показаны ссылки на видеофайл и исходный файл с расшифровкой, время начала предложения в виде отступа от начала видео в секундах и начало массива словоформ, составляющих предложение. Первое слово имеет единственный грамматический разбор, включающий поля *lex* (лемма), *transl_en* (перевод на английский язык) и поле *gr* (грамматика), которое, в свою очередь, распадается на несколько полей, соответствующих грамматическим категориям. Вторая словоформа (*linkse*) не имеет разбора, а третья является знаком пунктуации, что отражено в поле *type* (тип токена).

Очевидным преимуществом представленного способа хранения является то, что извлечённые из базы данных документы, соответствующие предложению, можно почти без изменений передавать в интерфейс, поскольку они являются json-объектами. Модификации, которым эти объекты подвергаются перед передачей в интерфейс, включают удаление избыточной информации (например, ссылка на исходный файл расшифровки не нужна при просмотре результатов пользователем) и подсветку словоформ, являющихся результатом запроса (что реализуется добавлением к словоформе атрибута *highlight*).

Запрос к такой базе данных тоже имеет вид json-объекта. В самом простом варианте запрос представляет собой словарь, где ключам, представляющим названия атрибутов, соответствуют их желаемые значения. Для выполнения более сложных запросов, включающих логические функции или регулярные выражения, используется специальный язык запросов MongoDB. Ниже приводится пример такого запроса:

```
{"$or": [{"ana.gr.case": "gen"},  
{"ana.gr.number": "pl"}], "ana.lex": {"$regex": "^z.*"}}
```

В приведённом запросе используются специальные обозначения "\$or", обозначающее дизъюнкцию, и "\$regex", обозначающее поиск по регулярному выражению. Весь запрос предназначен для поиска в коллекции словоформ и означает «найти все словоформы, грамматический разбор которых имеет тэг родительного падежа или тэг множественного числа и лемма которых начинается с буквы z».

Информация, которая вводится пользователем в поля поисковой формы, передаётся серверу в виде GET-запроса и преобразуется сервером в поисковый json-запрос описанного выше вида, после чего отправляется на обработку в поисковую систему. Ответ, полученный от поисковой системы, передаётся обратно в интерфейс в виде json-объекта.

Пользовательский интерфейс

Онлайн-интерфейс корпуса по существу представляет собой одну страницу, которая может выполнять функции задания поискового запроса и отображения результатов. Все операции, связанные с

передачей данных между сервером и клиентом, выполняются с помощью технологии Ajax, что позволяет изменять содержимое страницы (например, при задании нового запроса или переходе на следующую страницу выдачи) без её перезагрузки. Эта страница состоит из шапки и поисковых полей. В шапке находятся ссылки на информацию о проекте, инструкцию по использованию корпуса, а также на вход в панель администратора. Также там расположены кнопки для смены языка страницы.

В поисковых полях пользователь может задать лемму, словоформу (в том числе с помощью регулярных выражений), а также грамматические характеристики (либо с помощью чекбоксов, либо в виде строки, содержащей тэги и логические функции). Поисковое поле можно свернуть для удобного просмотра результатов выдачи.

Выдача подгружается на главную страницу под поисковым полем. С левой стороны расположен видеоплеер, в который загружаются необходимые видеофайлы. Из видеофайлов формируется плейлист, за счёт чего на странице находится только одно окно с видео, а не таблица из видео, как это сделано во многих других мультимедийных корпусах.

Для нашего корпуса мы использовали плеер Projekktor V1.3.09 (<http://www.projekktor.com/>). Этот плеер обладает рядом встроенных функций, которые показали нам полезными для нашего проекта, – формирование плейлиста и установка курсора на определённые фрагменты видео, необходимые для синхронизации с текстом.

Справа от видео расположен текст выдачи. Окно с текстом имеет такую же высоту, что и видеоплеер, и сопровождается полосой прокрутки. Полоса прокрутки, а также сворачивание поискового окна реализованы с помощью специального модуля, использующего функционал библиотеки jQuery.

В поисковой выдаче словоформы, являющиеся результатом поиска, выделены оранжевым цветом, при наведении курсора на каждое слово появляется морфологическая разметка в маленьком всплывающем окне. Во время воспроизведения видео синхронизированные фрагменты текста подсвечиваются.

ПЕРСПЕКТИВЫ

Одним из приоритетных направлений развития мультимедийного корпуса является увеличение количества текстов. С точки зрения разметки существует несколько возможностей для улучшения. Во-первых, в будущем может быть добавлен отдельный слой расшифровки с точной фонетической записью, который будет сопровождать основной слой с орфографической записью. Во-вторых, с помощью автоматических систем принудительного выравнивания (*forced alignment*) фрагменты, являющиеся единицами синхронизации, могут быть уменьшены до одного слова или даже до одной фонемы, что упростит фонетические исследования с помощью корпуса. Наконец, в-третьих, к грамматической разметке может быть добавлена другая разметка, релевантная для мультимедийных материалов, например, разметка жестов.

СПИСОК ЛИТЕРАТУРЫ

1. Кирьянов Д.П., Лучина Е.С., Панова Т.А., Тагабилева М.Г. Корпус языка Идиш: теория и практика // Тирош – труды по иудаике. – 2014. – № 14. – С. 78–90.
2. Гришина Е.А. Мультимедийный русский корпус (МУРКО): проблемы аннотации // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. – СПб.: Нестор-История, 2009. – С. 175–214.
3. MacWhinney B. From CHILDES to TalkBank // Research on Child Language Acquisition / eds. M. Almgren, A. Barreña, M. Ezeizaberrera, I. Idiazabal & B. MacWhinney. – Somerville, MA: Cascadia, 2001 – P. 17–34.
4. Anderson J., Beavan D., Kay C. SCOTS: The Scottish corpus of texts and speech // Creating and Digitizing Language Corpora / eds. J. Beal, K. Corrigan, H. Moisl. – 2007. – Vol. 1: Synchronic Databases. – Basingstoke: Palgrave Macmillan. – P. 17–34.
5. Neidle C., Vogler C. A New Web Interface to Facilitate Access to Corpora: Development of the ASLLRP Data Access Interface // The 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, LREC 2012. – Istanbul, Turkey, May 27, 2012.
6. Mayumi Bono, Kouhei Kikuchi, Paul Cibulka, Yutaka Osugi. A Colloquial Corpus of Japanese Sign Language: Linguistic Resources for Observing Sign Language Conversations // Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014. – Reykjavik, Iceland, May 26–31, 2014.
7. Kohn K. The BACKBONE project: pedagogic corpora for content and language integrated learning. Objectives, methodological approach and outcomes // Eurocall Review. – 2012. – Vol. 20, № 2.
8. Качинская И.Б. Корпус Диалектных Текстов в Национальном корпусе русского языка: состояние и перспективы // Лексический атлас русских народных говоров (Материалы и исследования). – СПб., 2009. – С. 57–68.
9. Летучий А.Б. Корпус диалектных текстов: задачи и проблемы // Национальный корпус русского языка: 2003–2005. – М.: Индрик, 2005. – С. 215–232.
10. von Waldenfels R., Daniel M., Dobrushina N. Why Standard Orthography? Building the Ustyia River Basin Corpus, an online corpus of a Russian dialect // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Беласово, 4–8 июня 2014 г.). Вып. 13 (20). – М.: Изд-во РГГУ, 2014.
11. Schaechter Mordkhe. Fun Folkshprakh tsu Kulturshprakh [The History of the Standardized Yiddish Spelling]. – New York: League for Yiddish/YIVO, 1999. – 111 с.
12. Muysken P. Bilingual speech: A typology of code-mixing. – Cambridge University Press, 2000. – 306 с.
13. Davies M. The advantage of using relational databases for large corpora: speed, advanced queries and unlimited annotation // International Journal of Corpus Linguistics. – 2005. – Vol. 10, № 3. – P. 307–334.
14. Даниэль М.А., Поляков А.Е., Рубаков С.В., Левонян Д.В., Плунгян В.А., Хуршудян В.Г. Восточноармянский национальный корпус // Армянский гуманитарный вестник. Вып. 2/3-II. – М.; Ереван: Зангак-97, 2009. – С. 9–33.
15. Evert S., Hardie A. Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. – 2011.
16. Аброскин А.А. Поиск по корпусу: проблемы и методы их решения // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. – СПб.: Нестор-История, 2009. – С. 277–282.
17. Sliwkanich T. et al. Towards scalable summarization and visualization of large text corpora // Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. – ACM, 2012. – P. 863.
18. Niekler A., Wiedemann G., Heyer G. Leipzig Corpus Miner – A Text Mining Infrastructure for Qualitative Data Analysis // Terminology and Knowledge Engineering 2014. – 2014.

Материал поступил в редакцию 12.12.14.

Сведения об авторах

АРХАНГЕЛЬСКИЙ Тимофей Александрович – кандидат филологических наук, преподаватель факультета филологии Национального исследовательского института «Высшая школа экономики», Москва
e-mail: tarkhangelskiy@hse.ru

СОЗИНОВА Ольга Андреевна – студент факультета филологии Национального исследовательского университета «Высшая школа экономики», Москва
e-mail: oa.sozinova@gmail.com

И.В. Якушевич

Знаково-семантическая структура поэтического символа и символа-симулякра

Описано знаково-семантическое структурирование символа и выявлены те изменения, которые он претерпевает, становясь симулякром в наименованиях фирм и предприятий.

Ключевые слова: символ, когнитивия, означаемое, означающее, денотат, симулякр, суггестивность

В современной науке отсутствует единая общепринятая теория символа, который на сегодняшний момент имеет столько же определений, сколько научных концепций включают его в свой метаязык. Структурная лингвистика скептически относится к символу как к объекту научного исследования в силу его междисциплинарного статуса: символ исследуется философами, культурологами, психологами, литературоведами и др.

Лингвистическая концепция символа реализуется в рамках когнитивно-семиологического метода исследования, корни которого восходят к теории Ф. де Соссюра о существовании слова в языке и речи, к трудам Э. Бенвениста о двойном означивании языковых единиц и позже – к работам И.А. Бодуэна де Куртенэ о билатеральной сущности языка. В наше время когнитивно-семиологический метод описан в трудах Ю.С. Степанова [1] и его последователей: А.А. Уфимцевой [2], Н.Н. Болдырева [3], Н.Ф. Алефиренко [4]. За рубежом направление разрабатывалось Р. Бартом [5], А.-Ж. Греймасом [6], Ж. Делезом [7].

Сущность этого метода заключается в описании слова на пересечении нескольких «измерений»: **когнитивии, знака, семантики и дискурса** [6, с. 104]. В центре этого пересечения мы помещаем и слово-символ. Среди постулатов когнитивно-семиологического метода Н.Ф. Алефиренко для нашего исследования важны следующие:

Во-первых, **символ является разновидностью когнитивии**. Это континуальное единство самого познавательного процесса (получение информации, ее обработка), результата (некоего знания), и материализации этого знания в языке и речи [8]. Таким образом, сама суть когнитивии обуславливает языковую природу символа.

Во-вторых, **семантику символа следует рассматривать как в языке, так и в дискурсе** [3, с. 99].

В-третьих, в речи **семантика слова-символа многослойна**. А.-Ж. Греймас выделил два уровня речи – семиологический и семантический [6, с. 78]: на сформированную в контексте культуры семантику символа наслаивается дискурсивная семантика, обес-

печивая «двухуровневое существование значения» [4, с. 105].

В четвертых, **слово-символ – это знак**. Однако, если у классического знака (например, слова) означающее – это только его фонетическая материальная оболочка, то у символа означающим является все слово в единстве его фонемного ряда и стоящего за ним чувственного образа. В свою очередь, слово-означающее само может подвергаться символизации и указывает на специфическое для символа означаемое – мифологические представления о мироустройстве и человеческом бытии: ‘мир’, ‘бог’, ‘жизнь’, ‘смерть’, ‘душа’ и пр. [5, с. 79].

К примеру, **означающее** А символа АНГЕЛ (лексема «ангел») состоит из фонемного ряда < а́нг’эл > и прямого понятийного значения, создающего эмпирическое представление о денотате: «сверхъестественное существо, посланец Бога (изображается обычно в виде крылатого юноши). По небу полуночи ангел летел» [9].

Означаемое В состоит из девяти символических значений, выявленных нами на основании повторяющихся данных двух словарей символов [10, с. 41–45; 11, с. 41–45; 12, с. 12–13]: 1) воля Бога, 2) воздух, 3) дух, 4) бессмертие, 5) огонь, 6) защита, 7) красота, добро, 8) смерть, 9) рождение, жизнь (рис. 1).

Означаемое слова-символа уникально, так как:

1) многозначно и незамкнуто, поскольку находится в процессе формирования (П. Рикер, Ж. Лакан, А.Ф. Лосев, Ю.М. Лотман);

2) содержит метафизическую космо- и антропоцентричную семантику (мироздание, космос, бытие, природа, жизнь и смерть, здоровье и пр.);

3) инвариантно в контексте культуры (К. Леви-Стросс, И. В. Сафьянова и др.);

4) является элементом мифа (А.А. Потебня, А.Н. Афанасьев, О.Н. Трубачев, В.В. Колесов и др.);

5) обладает суггестивностью, то есть регулярной воспроизводимостью вне контекста (К. Леви-Стросс, Ж. Делез, А.Н. Веселовский).

Для понимания механизма символизации и превращения символа в симулякр важно учитывать зна-

чимость денотата – предмета или целого фрагмента действительности, на которое указывает слово-означающее [13, с. 51]. Для наглядности обратимся к классическому треугольнику Огдена-Ричардса. Как билатеральный знак символ имеет два типа значения: чувственный образ означающего и мифологическую абстрактность означаемого. Следовательно, и денотатов у него тоже минимум два. А вот материальный носитель – фонемный ряд – один. Покажем эту особенность на рис. 2.

На схеме видно, что треугольники означающего и означаемого подобны. В отношении подобия находятся и их углы: верхние (денотаты) и правые (зна-

чения, указывающие на денотат). Следовательно, за каждым словом-символом стоит минимум два подобных денотата и два значения, отражающих это подобие. В этом заключается главное уникальное свойство символа – **полиденотативность**. И означающее, и означаемое символа соотнесены каждый со своим денотатом (в дискурсе – с референтом), которые сосуществуют параллельно как два подобных фрагмента действительности, каждый со своим хронотопом. Полиденотативность символа обуславливает **мифологическое мышление двоемирия**, когда сознание говорящего раздвоено и присутствует как в реальности означающего, так и означаемого.

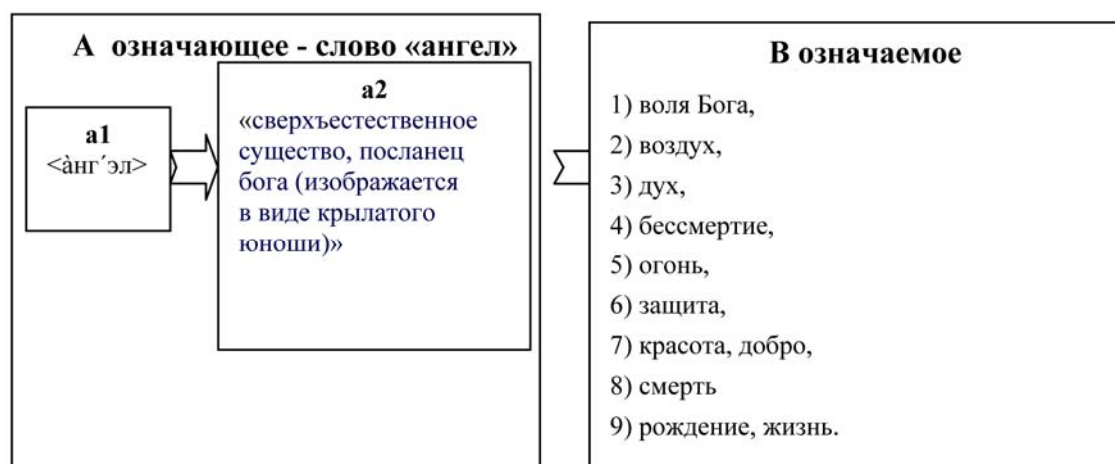


Рис.1. Лексико-семантическая структура лексемы-символа «ангел»

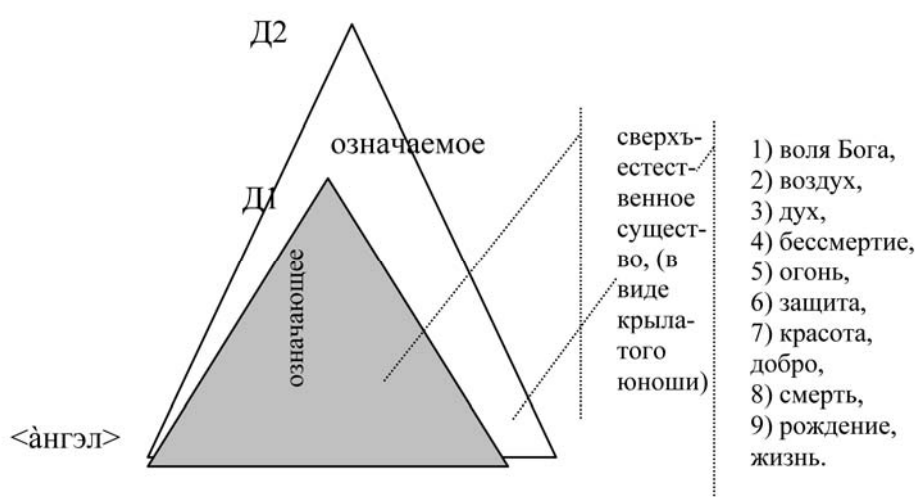


Рис. 2. Символ в треугольнике Огдена-Ричардса

Двоемирие символа особенно важно в поэзии для материализации переживаний лирического героя или осознания судьбоносных для всякого человека событий – любовь, рождение, смерть. Так, в стихотворении Арсения Тарковского «Четвертая палата» означающее символа АНГЕЛ→СМЕРТЬ выражено словосочетанием *больничный ангел*.

*Ангел больничный за шторой
Светит одеждой туманной.
— Я за больной.
— За которой?
— Я за детдомовской Анной.*

<...>

*Востренький нос, восковые
Пальцы, льняная косица.
Мимо проходят живые.
— Что тебе, Анька?
— Не спится.*

Если в языке символ, сформированный в культуре народа, является максимально обобщенным **инвариантом**, или символом-прототипом, то в дискурсе символ-прототип вербализован в тех или иных **языковых вариантах** [8, с. 104]. В приведенном отрывке означающее выражено словом *ангел*. Означающее ‘смерть’ – словами, ассоциативно связанными с семантикой смерти и тем самым косвенно ее манифестирующими: *востренький нос, восковые пальцы*. Фраза ангела *я за больной*, напоминает о поверии, будто ангелы-хранители уносят на небеса душу умершего праведника. Реальность сидящего на окне ангела и реальность умирающей девочки сосуществуют.

В отрывке вербализованы оба компонента символа. Но в дискурсе часто встречаются **одночленные символы**, которые мы называем так по аналогии с одночленными метафорами [14]. В одночленном символе, как и метафоре, не вербализовано означаемое. Однако его символические значения в сознании так или иначе актуализируются, независимо от того, что ближайший контекст их не поддерживает. Это явление называется **суггестивностью** – способностью означающего символа вызывать ряд символических значений, невербализованных, но прочно закрепленных за этим образом в культуре народа и на этом основании возникающих в сознании.

В лирике одночленные символы частотны. Вне литературы мы встречаемся с ними в функции названий торговых марок, общественных предприятий и фирм: «Корона» (прокат лимузинов), «Абрикосовый лев» (агентство аниматоров), «Домашний очаг» (сеть магазинов по реализации предметов интерьера) и др. Рассмотрим семантическую структуру названия московского автомобильного клуба по оказанию технической помощи на дороге и вызову эвакуатора «Ангел». Важно определить, как в функции названия изменилась вся семантическая структура слова-символа и насколько актуализировано его символическое означаемое. В этой функции происходит разрыв естественно сложившейся связи между фонетической оболочкой <анг’эл> и всей системой значений слова-символа. Место означающего-чувственного образа занимают понятия: ‘автомобильный клуб’,

‘ремонт’, ‘эвакуатор’. Происходит **переадресация к новому референту**. Вместо многомирия символа с двумя планами референции мы получаем слово с одной ложной референцией.

Именно раскол между означающим и означаемым и, как следствие, переадресация к ложному референту и есть, по мнению Ж. Бодрийяра, главные признаки симулякра. Истинный, но отвергнутый референт, ангел, «находится вне знака не в большей степени, нежели означаемое... у него нет никакой реальности – кроме той, что вписывается между строк знаков». [15, с. 185, 186]. Он становится потенциальной фоновой семой в лексическом значении названия. Не могут исчезнуть и символические значения, которые по воле свойственной символу суггестивности, независимо ни от какого контекста и даже будучи невербализованными все равно возникают в сознании: ангел→воля Бога, ангел→дух, ангел→защита и пр. В семантической структуре названия фирмы они также становятся потенциальными фоновыми семами. Семантическая структура названия фирмы «Ангел» такова: **А означающее:** <анг’эл>→ **В означаемое:** 1) понятийное значение: ‘автомобильный клуб’, ‘ремонт’, ‘эвакуатор’; 2) фоновые потенциальные семы: ‘сверхъестественное существо’, ‘посланец Бога (изображается в виде крылатого юноши)’, ‘воля Бога’, ‘воздух’, ‘дух’, ‘бессмертие’, ‘огонь’, ‘защита’, ‘красота’, ‘добро’, ‘смерть’. Некоторые фоновые семы актуализируются в дискурсе интернет-сайта фирмы. Так символическое значение ‘защита’ читается в словах и предложениях *друг, помощь, он приходит в трудную минуту, чтобы помочь; автоангел – это больше, чем просто ангел*. Кроме того, сайт фирмы украшен изображением крыла.

Впервые теория симулякра возникает в диалоге Платона «Софист», где он выделяет два вида искусства: искусство творить образы (когда автор образов передает «истинную соразмерность прекрасных вещей») и искусство создавать призрачные подобию (призраки) [16, с. 299–300], когда «истинная сущность» вещей отдалается или совсем стирается: «о несуществующем говорят как о существующем» и «о существующем – не так, как оно есть» [там же, с. 338, 340]. Это платоновское подобие и есть симулякр. Делез возрождает это понятие для описания эстетики и художественной литературы и заостряет проблему, которая ложится в основу теории симулякров Бодрийяра и современных исследований (И.А. Мальковская, С.В. Лещев): «властью симулякров определяется современность» [7, с. 346].

Одна из причин потребности современного общества в симулякре – его мощная манипулятивная способность. Для Бодрийяра симулякр и манипуляция – явления неразрывные [15, с. 185, 186]. Манипуляция – это «вид психологического воздействия, искусное исполнение которого ведет к скрытому возбуждению у человека намерений, не совпадающих с его актуально существующими желаниями» [17, с. 52]. Переадресация всех символических значений другому референту наделяет его не просто новыми фоновыми семами, а коннотативным значением, заданно положительным или отрицательным в соответствии с желаниями и целями адресанта. Таким образом, в на-

званиях фирм содержится скрытая реклама. К примеру, фоновые значения 'ангел', 'красота', 'добро', 'воля Бога' и др. формируют положительную коннотацию фирме, предоставляющей эвакуатор и помощь на дорогах, выделяет ее из ряда других компаний и, в конечном счете, рекламирует ее. А если учесть, что символ формируется в культуре народа и не только создается индивидуум, но и хранится в безднах его подсознания, то сила символического внушения велика и способствует успешной манипуляции.

Приведем еще пример. Научно-производственный центр «Огонек» является одним из ведущих российских предприятий реабилитационной индустрии и специализируется на производстве ортопедических изделий, организации восстановительного лечения больных с поражением нервной системы и опорно-двигательного аппарата, разработке медицинской компьютерной техники для реабилитации и скрининг-диагностики и пр. В основе названия центра – одна из моделей символа «огонь»: ОГОНЬ→ЖИЗНЬ, ЗДОРОВЬЕ [18, с. 137; 19, с. 337]. О модели ОГОНЬ→ЖИЗНЬ читаем у Афанасьева: «Если душа понималась как **огонь**, то **жизнь** возможна была только до тех пор, пока горело это внутреннее пламя; погасало оно – и жизнь прекращалась. У нас уцелело выражение: погасла жизнь; выражение это в народной песне заменено сравнением смерти человека с погасшею свечою» [11, с. 199]. Кроме того, эта модель легла в основу устойчивого словосочетания *живой огонь*: «самородный, лесной, деревянный, царский, вытертый из дерева; это делается на Ивана Купала» [20, с. 644]. В словаре К.С. Горбачевича *жизнь* имеет эпитеты *горячая, остывшая (уст.)* [21, с. 170]. Если *огонь* – символизирует жизнь и здоровье взрослого человека, то *огонек* – маленького человека, т.е. ребенка. Действительно, центр «Огонек» преимущественно специализируется на современных методиках лечения детского церебрального паралича, последствий черепно-мозговых травм и врожденных аномалий центральной нервной системы. Это символическое значение стало фоновой семой в многослойной семантике симулякра «огонек»: **А означающее:** <огон'ок>→**В означаемое:** 1) понятийное значение 'центр реабилитационной медицины прежде всего для детей с ДЦП'; 2) фоновые потенциальные семы: 'маленький огонь', 'жизнь', 'ребенок', 'здоровье'. В одной из статей о центре некоторые фоновые значения актуализированы: «Если диагноз ДЦП появился в медицинской карте вашего **ребенка** и разделил вашу **жизнь** на „до“ и „после“, если вы уже прошли стадию отчаяния и безысходности и приняли решение не опускать руки, вы неизбежно приходите к пониманию того, что успех развития вашего **ребенка** в основном определяется систематическим воспитанием активных движений» [22]. В структуре обычного символа 'огонь' как 'свещающиеся газы' [9] – понятийное означающее, а в симулякре – только фоновое значение с положительной коннотацией, нашедшее отражение в логотипе компании: первая буква «о» изображена в виде двух языков пламени.

Рассмотрим пример, на наш взгляд, не совсем удачного символа-симулякра: московская компания по вывозу мусора «Феникс». В основе названия – символическая модель ФЕНИКС→ОГОНЬ и ФЕНИКС→БЕССМЕРТИЕ [12, с. 389; 23, с. 638, 639]. В мифологии некоторых древних народов Феникс – сказочная птица, в старости сжигаящая себя и вновь возрождающаяся из пепла молодой (употребляется как символ вечного обновления, возрождения) [9]. В слове-символе это означающее, в символе-симулякре – только фоновая семантика, актуализированная в логотипе фирмы: изображение оранжевой птицы, у которой вместо крыльев – языки пламени. Означающее симулякра <ф'эн'икс>. Означающее: 1) понятийное значение: 'фирма по вывозу и утилизации мусора'; 2) фоновые семы: 'птица', 'огонь', 'пламя', 'возрождение'. Сема 'огонь', вероятно, подчеркивает тот факт, что «Феникс» не только вывозит, но и утилизирует мусор, что, безусловно, повышает его рейтинг. Это тем более важно, что в начале 2015 г. состоится московский конкурс отбор, на котором будут выбраны инвесторы, способные создать не только универсальные полигоны, но и комплексные организации по вывозу и переработке бытового и строительного мусора.

Однако вернемся к семантической структуре символа-симулякра «феникс». Поскольку механизм символизации предусматривает подобие означающего и означаемого, фоновая сема 'возрождение' должна быть соотнесена с семой 'мусор' из понятийного значения. В результате возникает нежелательное для имиджа предприятия значение возрождения мусора, которое порождает сомнения в том, действительно ли осуществляется утилизация, насколько она эффективна, не продуцирует ли она дополнительный мусор и т.д.

Сделаем выводы:

1. Символ в когнитивно-семиологическом аспекте – это лексема со сложной многослойной структурой и несколькими подобными друг другу денотатами, существующими одновременно в параллельных фрагментах действительности. Его означающее – чувственный образ, названный словом, а означаемое – система константных в культуре народа и мифологии значений.
2. Символ как название фирмы или предприятия становится симулякром, поскольку все значения символа – и понятийные и символические – переадресованы другому референту и в структуре его лексического значения становятся фоновыми.
3. Символ-симулякр – эффективный манипулятивный механизм, поскольку переадресованные фоновые семы бывшего символа создают положительную коннотацию новому референту и тем самым рекламируют его, подсознательно заставляя клиентов выбирать именно эту фирму и пользоваться ее услугами.

СПИСОК ЛИТЕРАТУРЫ

1. Степанов Ю.С., Эдельман Д.И. Семиологический принцип описания языка // Принципы описания языков мира. – М.: Наука, 1976. – С. 247–281.

2. Уфимцева А.А. Лексическое значение: Принципы семиологического описания лексики. – М.: Наука, 1986. – 240 с.
3. Болдырев Н.Н. Когнитивная семантика. – 3-е изд. – Тамбов: Изд-во Тамб. ун-та, 2002. – 172 с.
4. Алефиренко Н.Ф. Когнитивно-семиологическая теория слова // Вестник СамГУ. – 2006. – № 5/1. – С. 102–110.
5. Барт Р. Миф сегодня // Избранные работы: Семиотика. Поэтика. – М.: Изд-во им. Сабашниковых, 2004. С. 233–286.
6. Греймас А.-Ж. Структурная семантика: поиск метода. – М.: Академический проект, 2004. – 368 с.
7. Делез Ж. Логика смысла. – М.: Раритет, 1998. – 480 с.
8. Болдырев Н.Н. Языковые механизмы оценочной категоризации // Реальность, язык и сознание: Междунар. межвуз. сб. науч. тр. – Тамбов, 2002. – С. 360–369.
9. Словарь русского языка: в 4 т. / под ред. А.П. Евгеньевой. – 4-е изд. – М.: Русский язык; Полиграфресурсы, 1999. – URL: <http://feb-web.ru/feb/mas/mas-abc>.
10. Ангел. Энциклопедия символов, знаков, эмблем / сост. В. Андреева, В. Куклев, А.М. Ровнер. – М.: Локид; Миф, 1999. – 576 с.
11. Афанасьев А. Н. Поэтические воззрения славян на природу: в 3 т. – М: Индрик, 1994. III. Т. III. – 784 с.
12. Трессидер Дж. Словарь символов. – М.: ФАИР-ПРЕСС, 2001. – 448 с.
13. Солодуб Ю.П. Альбрехт Ф.Б. Современный русский язык. Лексика и фразеология (сопоставительный аспект). – М: Флинта, Наука, 2002. – 264 с.
14. Солодуб Ю. П. Структурная типология метафоры // Филологические науки. – 1999. – № 4. – С. 67–76.
15. Бодрийяр Ж. К критике политической экономики знака. – 2-е изд. – М.: Библион-Русская книга, 2004. – 304 с.
16. Платон. Федон, Пир, Федр, Парменид / под общ. ред. А.Ф. Лосева, В.Ф. Асмуса, А.А. Тахо-Годи. – М.: Мысль, 1999. – 528 с.
17. Доценко Е. Л. Психология манипуляции: феномены, механизмы и защита. – М.: ЧеРо, изд-во МГУ, 2005. – 344 с.
18. Копалинский В. Огонь // В. Копалинский. Словарь символов. – Калининград: ФГУИПП «Янтарный сказ», 2002. С. 136–140.
19. Огонь Славянская мифология. Энциклопедический словарь. – М.: Международные отношения, 2001. С. 22–23.
20. Даль В.И. Толковый словарь живого великорусского языка: в 4 т. – М.: ТЕРРА, 1994. Т. II. – 784 с.
21. Горбачевич К. С. Словарь эпитетов русского литературного языка. – СПб.: Норинт, 2001. – 221 с.
22. Титаренко Н.Ю. «Огонек» вчера, сегодня, завтра. Жизнь с ДЦП // Информационно-практический журнал. – URL: http://www.cplife.ru/n11/articles5_1.html.
23. Феникс // Энциклопедия сверхъестественных существ. – М.: Эксмо; СПб.: Мидград, 2006. – С. 638–640.

Материал поступил в редакцию 02.12.14.

Сведения об авторе

ЯКУШЕВИЧ Ирина Викторовна – кандидат филологических наук, доцент кафедры русского языка Московского государственного педагогического университета, Института гуманитарных наук
e-mail: sal107@yandex.ru

В.И. Хайруллин

Культурно-прагматическая адаптация при переводе как средство извлечения информации

Рассматривается семантико-терминологическое поле общественного института образования. Отмечается релевантность культурно-прагматической адаптации при переводе с английского языка на русский, способствующая извлечению информации, заложенной в англоязычных терминологических единицах. Указываются причины, по которым переводчик может отказаться от адаптации.

Ключевые слова: маркер, культура, адаптация, перевод, информация, признак.

В настоящее время, как указывают авторы [1], роль информации чрезвычайно возросла, что в определенном смысле делает необходимым уточнить некоторые понятия, имеющие хождение в науке о языке. Одним из наиболее актуальных в этом смысле представляется понятие и термин «образование».

В широком понимании термин и понятие «образование» означают «совокупность знаний, полученных в результате обучения» [2, с. 436]. Институт образования предполагает также понятие общей образованности, т. е. тех фоновых знаний, которыми располагает автор текста и/или реципиент как представители той или иной культуры.

В каждой культуре имеются маркеры – признаки, позволяющие идентифицировать описываемую ситуацию, – относящиеся к области общеобразовательных знаний, которыми предположительно владеет «средний» представитель культуры. Эти общие знания могут относиться к области художественной литературы, искусства, кинематографа и др.

В настоящей статье в качестве исследовательского материала используются англоязычные тексты, описывающие ситуации, относящиеся к сфере общеобразовательных знаний, и их переводы на русский язык. Методом исследования является метод сопоставления переводов, позволяющий наиболее ярко представить искомые особенности.

Анализ материала показал, что общеобразовательные маркеры культуры относятся к одним из наиболее распространенных в английском тексте, причем в ряде случаев они могут разъясняться непосредственно в тексте перевода:

I saw it against the background of Dufton, the back-to-back houses ..., the smoke which caught the throat and dirtied clean linen in a couple of hours, the sense of being always involved in a charade upon *Hard Times* [3, p. 35].

Перед моими глазами еще маячил Дафтон: сросшиеся друг с другом домишки..., дым, от которого вечно першит в горле и чистая рубашка через два-три часа становится черной, и, наконец, никогда

не покидающее тебя ощущение, что все это вроде инсценировки романа Диккенса «Тяжелые времена», в котором ты обречен принимать участие [4, с. 24].

С переводоведческой точки зрения в подобных примерах используется прием детализации, или экспликации, при котором лексические единицы исходного языка заменяются словосочетаниями, эксплицирующими значения первых, т. е. дающими более полное объяснение их значений на языке перевода [5, с. 251]. Однако такой прием используется далеко не всегда: значительно более частотны случаи, при которых англокультурные маркеры эксплицируются в тексте комментариев, сопровождающих как вариант перевода, так и вариант исходного языка, если англоязычный вариант заведомо рассчитан на русскоязычного реципиента, т. е. если англоязычное произведение издано в России:

Against my will I felt pleased that he should have considered my remarks interesting, though I knew that it was *Dale Carnegie* stuff, a small apparently casual compliment [3, p. 77].

Отметим, что на с. 77 московского издания книги дается следующий комментарий: «Dale Carnegie ... – американский лектор, писатель, очеркист. Автор книги «Как приобретать друзей и влиять на окружающих» (How to Win Friends and Influence People – 1936 г.).

Перевод на русский язык –

Я был невольно польщен тем, что он находит мои замечания интересными, хоть и понимал, что это было сказано неискренне – всего-навсего дешевый комплимент по рецепту Дейла Карнеги [4, с. 70] –

сопровождается примечанием редактора на с. 70: «Автор книги «Как приобретать друзей и влиять на окружающих».

В следующем примере название романа также комментируется как в английском, так и в русском варианте:

I read a lot of classical books, like *The Return of the Native* and all ... [6, с. 42]

Конечно, я читаю всякие классические книги вроде «Возвращение на родину» ... [7, с. 36].

(С. 223 англоязычного московского издания: «Возвращение на родину» (1878), роман английского писателя Томаса Гарди (Thomas Hardy, 1840-1928). С. 36 перевода на русский язык: «Возвращение на родину» – роман Томаса Гарди).

Те случаи, когда элементы англокультурного фонда в переводе не уточняются и не комментируются, представляют собой примеры «игнорирования» русскокультурного реципиента. В неадаптированных и некомментируемых переводах мы наблюдаем потерю той части информации, которую несут в себе англокультурные маркеры, и которая именно в силу монокультурного характера не может быть адекватно воспринята русскокультурным реципиентом, например:

Antheil's my man [8, p. 39]

Лично я – поклонница Антейля [9, с. 11]

(см. с. 347 англоязычного издания с русскими комментариями: «Antheil – Джордж Антейль (1900-1959), американский композитор-авангардист»).

Перевод на русский язык опубликован без комментариев).

Роль литературы и искусства в плане пополнения культурного фонда языка невозможно переоценить. Эти области являются постоянным источником признаков, которые могут быть широко применимы и понятны представителям исходной культуры, однако, для представителя целевой культуры они оказываются малоинформативными. По этой причине перевод нередко сопровождается «раскрытием», экспликацией инокультурного признака:

In 1933 he sent emissaries to the man who controlled all the gambling activities of Manhattan, the crap games on the docks, the *shylocking* that went with it... [10, p. 220-221].

В 1933 году он послал доверенных лиц к человеку, который управлял всем игорным бизнесом Манхэттана, будь то игра в кости на прибрежных улочках или *ростовщичество*, непременно сопутствующее ей... [11, с. 157].

В английском высказывании используется признак *shylocking*, происходящий от имени ростовщика Shylock'a (Шейлока) из комедии В. Шекспира «Венецианский купец». В английской культуре имя стало нарицательным: «*shylock n., a usurious money-lender*» [12, p. 487], т. е. «ростовщик, меняла». В русской культуре этот признак не имеет такого широкого хождения, как в английской. Поэтому перевод сопровождается экспликацией, а именно, использованием признака, поясняющего, что имеется в виду под термином *shylocking*, – «ростовщичество». Такой вид культурно-прагматической адаптации основан на замене персонифицированного признака исходного варианта обобщающим эксплицирующим признаком переводного варианта.

Культурно-прагматическая адаптация может быть основана не только на экспликации признака (маркера), но и на «скрадывании» инокультурного маркера в переводе, т. е. на его переводческом игнорировании или на замене исходного признака другим, основанном на приеме смыслового развития, при котором

значение лексической единицы переводящего языка является логическим следствием значения исходной единицы (единиц) [5, с. 248]. Использование данного приема представляется оправданным в тех случаях, когда маркеры не являются широко известными в целевой культуре, как например, имена британских издателей в следующем английском высказывании:

It snowed in Warley the night I left, just a light powdering, a present from *Raphael's* and *Tuck's* and *Sharpe's*... [3, p. 99].

В ночь, когда я покидал Уорли, выпал снег. Он лишь слегка припорошил улицы – это был как *новогодний* подарок ... [4, с. 93].

В данном английском примере сочетание признаков *snow*, *Raphael*, *Tuck*, *Sharpe* дало переводчику возможность обратиться к приему смыслового развития: коль скоро речь идет о снеге и зиме и при этом упоминаются имена издателей, выпускающих в частности, рождественские и новогодние открытки, то вполне уместным представляется использование признака «рождественский» или «новогодний», который значительно более близок русскокультурному реципиенту, нежели *Raphael*, *Tuck* и *Sharpe*.

Перечисленные особенности не исчерпывают всей специфики культурно-прагматической адаптации. Важное место в системе общих культурных ценностей занимает кинематограф, дающий ценный исследовательский материал.

Как отмечалось, при переводе английского высказывания на русский язык может использоваться уточняющий признак, который эксплицирует культурные маркеры:

She and Mrs Fosnacht have gone into Brewer to a matinee of *Some Like It Hot* [13, p. 126].

Она с миссис Фоснахт уехала в Бруэр на утренний сеанс *фильма* «Некоторые любят погорячее» [14, с. 126].

Если для англокультурного реципиента информативным является название *Some Like It Hot*, без указания того, что речь идет о фильме, поскольку эта картина, как и многие другие с участием Мэрилин Монро, пользовалась огромным успехом, то для русского реципиента название «Некоторые любят погорячее» малоинформативно, он может лишь предполагать, что это название какого-либо зрелищного мероприятия – спектакля или, может быть, фильма. Для устранения этой неопределенности переводчик адаптирует маркер, эксплицируя, что речь идет именно о фильме, тем более что в российском кинопрокате эта картина известна под названием «В джазе только девушки».

В силу некоторых культурно-социальных причин название ежегодной премии Американской академии киноискусства «Оскар» хорошо известно в России. Возможно, это связано с тем, что время от времени премия присуждалась и советским фильмам как лучшим иностранным фильмам года. Так или иначе, признак «Оскар» вызывает реакцию у русскокультурного реципиента, он ему знаком лучше, чем академия, присуждающая премию с этим названием. Вероятно, этим объясняется использование маркера

«Оскар» в русских переводах вместо маркера *Academy Award* исходных английских вариантов:

Now the Godfather was hurt, in trouble, and he could kiss the Academy Award goodbye... [10, p. 192].

Но теперь крестный отец сам пострадал, попал в беду, ...так что прости-прощай заветный «Оскар» [11, с. 135].

По нашим наблюдениям, признак «Оскар» более знаком в русской культуре, чем признак «премия академии» (*Academy Award*). В подтверждение приведем в качестве примера текст под фотографией, помещенной в газете *The Moscow Tribune*, рассчитанной, прежде всего, на русского читателя. Фотография запечатлела момент вручения «Оскара» С. Спилбергу, режиссеру фильма «Список Шиндлера». В этом тексте четко указаны как название премии, так и название мероприятия, посвященного вручению этой премии:

Harrison Ford, right, presents Steven Spielberg the Oscar for best Director for Schindler's List at the 66th Academy Award in Los Angeles [15, p. 4].

А вот текст под фотографией из газеты *International Herald Tribune*, рассчитанной на западного читателя. В тексте название премии не уточняется:

Spielberg won awards for best film and best director... [16, p. 24].

Вместе с тем, в англоязычной культуре есть маркеры, которые нерелевантны для русского читателя. В переводе такие маркеры либо игнорируются («скрадываются»), либо заменяются русскокультурными маркерами, например:

It took him ten minutes to find the *Encyclopedia Britannica* [17, p. 199].

Ему потребовалось целых десять минут, чтобы разыскать нужный том *энциклопедии* [18, с. 126].

В англоязычном мире Британская энциклопедия является наиболее авторитетной, поэтому информация о том, что он искал именно Британскую энциклопедию, представляется ценной для англокультурного читателя. Для русского читателя ценна информация, что он искал энциклопедию. Читатель может не знать, что Британская энциклопедия – это издание, рассчитанное не только на Британию, но и на другие страны, поэтому информация о том, что кто-то ищет Британскую энциклопедию, хотя действие происходит в США, может вызвать недоумение. Мы предполагаем, что это одна из тех причин, по которым русский перевод «скрадывает» культурологическую информацию английского варианта.

Однако в некоторых случаях культурная адаптация при переводе представляется излишней, например:

Let's have the file, Doc, from A to Z... [19, p. 50].

Давайте сюда все ящики от А до Я... [20, с. 14].

В данном случае мы полагаем, что маркер «от А до Z» можно было бы сохранить в русском переводе, поскольку каждый читающий по-русски сознает, что английский алфавит основан на латинском, и в нем не может быть буквы «я». Излишняя адаптация иногда приводит к курьезам, как в данном случае. Хотя, с другой стороны, этот пример можно интерпретиро-

вать как стремление переводчика учесть тот факт, что каждый реципиент перевода находится в плену стереотипов своей культуры, он прежде всего проецирует модели чужой культуры на модели своей собственной. Без такой проекции некоторые инокультурные признаки в ряде случаев бывают непонятны. Интересный пример в этой связи приводит американский исследователь. Как-то раз во время поездки в Югославию он проходил мимо настенного термометра, который показывал 50 градусов. Эти показания ему ничего не говорили и он на них никак не отреагировал, пока по таблице преобразования температур не определил, что 50 градусов по Цельсию составляют 122 градуса по Фаренгейту. Только тогда он понял, что было действительно жарко [21, p. 376].

Что означает признак «от А до Я» в нашем примере? Он означает весь алфавит – от первой до последней буквы. Если бы в переводе был указан признак «от А до З», это выглядело бы странно, поскольку в русском алфавите «З» не последняя, а девятая буква!

Культурно-прагматическая адаптация необходима в таких случаях, которые на лингвистическом уровне можно было бы назвать случаями «ложных друзей» переводчика, например:

He came from St Louis, he told her ..., and had finished *high school* [17, p. 36].

Арнольд родился в Сент-Луисе, там же окончил *среднюю школу* [18, с. 38].

На культурном уровне термин *high school* не соответствует тому, что по-русски называется «высшая школа», поскольку «высшая школа» – это университет, т. е. учебное заведение, дающее высшее образование. В англоамериканской культуре *high school* – это старшие классы средней школы, в отличие от *low* (или *lower*) *school*:

You always got these very lumpy mashed potatoes on steak night, and for dessert you got Brown Betty, which nobody ate, except maybe the little kids in the *lower school* that didn't know any better [6, p. 57].

А в англобританской культуре *public school* – это отнюдь не «общественная», а частная школа, система образования в которой напоминает скорее университет, информация о чем может присутствовать в тексте перевода:

I was remembering ... his genuine *Public School* accent [3, p. 162].

Я вспомнил ... его *университетское* произношение [4, с. 159].

В ряде случаев англокультурные маркеры подвергаются значительной адаптации, поскольку некоторые из них весьма дистанцированы от русской культуры, и поэтому при переводе бывает трудно подобрать соответствия, например, таким признакам, как Oxford drawl, Standard English и др. В переводе на русский язык такие признаки передаются «приблизительно», не совсем точно, так как целевая культура, с одной стороны, не располагает сходными им признаками, с другой, они не могут быть представлены в том виде, как в исходном тексте, потому что связь между ними и английской культурой нарушается,

поскольку перевод рассчитан на русскокультурного реципиента:

He ... spoke English with a slow *Oxford drawl* [8, p. 35].

... По-английски он говорил, *слегка растягивая слова на оксфордский манер* [9, с. 7].

To speak *Standard English* is in itself suspect ... [3, p. 150].

Собственно, если человек говорит на *очень правильном языке*, на него уже смотрят с подозрением ... [4, с. 147].

I was going to pronounce *Aunt with a broad a* ... [3, p. 27].

Я хотел было произнести «Эмили» на *оксфордский манер*, ... [4, с. 15].

Известно, какое большое значение имеет правильное произношение в английской культуре, хотя этим «правильным» английским владеет всего-навсего три процента населения Соединенного Королевства. Приводимые нами примеры иллюстрируют эту особенность английской культуры. Правильное произношение (Standard English; Received Pronunciation; BBC English) служит индикатором высокого социального происхождения, что хотели подчеркнуть, например, голливудские режиссеры, заставляя актеров в своих фильмах имитировать высокосоциальный британский акцент, а директора крупных американских компаний нанимали секретарш из числа образованных англичанок, чтобы придать своему бизнесу хороший тон [22, p. 28].

Термины *Standard English*, *Received Pronunciation*, *BBC English*, *Oxford drawl* неизвестны в русской культуре, поэтому в русском языке они передаются с использованием приема описательного перевода – «очень правильный язык», «(на) оксфордский манер» и т.д.

Проведенное исследование позволяет отметить, что метод сопоставления переводов обладает большим эвристическим потенциалом, что делает целесообразным использовать его как в терминоведческой деятельности [23, с. 30-32], так и при осуществлении целого ряда других важных научных проектов.

СПИСОК ЛИТЕРАТУРЫ

1. Караваев Н.Л. Информационное общество: попытка осмысления сущности понятия // Научно-техническая информация. Сер. 1. – 2014. – № 6. – С. 1-5.
2. Ожегов С.И., Шведова Н.Ю. Толковый словарь русского языка. – М.: А ТЕМП, 2010. – 874 с.
3. Braine J. Room at the Top. – М.: Foreign Languages Publishing House, 1961. – 272 p.
4. Брэйн Дж. Путь наверх / пер. с англ. Т. Озерской и Т. Кудрявцевой. – М.: Изд-во иностр. литературы, 1960. – 263 с.
5. Комиссаров В.Н. Теория перевода. – М.: Высшая школа, 1990. – 253 с.
6. Salinger J.D. The Catcher in the Rye. – М.: Progress, 1979. – 248 p.
7. Сэлинджер Дж. Над пропастью во ржи // Сэлинджер Дж. Над пропастью во ржи: Повести. Рассказы / пер. Р. Райт-Ковалевой. – М.: Правда, 1991. – С. 21-184.
8. Fitzgerald F.S. Tender is the Night. – М.: Raduga, 1983. – 400 p.
9. Фицджеральд Ф.С. Ночь нежна / пер. Е. Калашниковой. – М.: Панорама, 1993. – 304 с.
10. Puzo M. The Godfather. – London-Sydney: Pan Books, 1974. – 448 p.
11. Пьюзо М. Крестный отец // Пьюзо М. Крестный отец. Сицилиец. – Баку: Олимп, 1991. – С. 3-296.
12. The New American College Dictionary. – New York: Signet Book, 1981. – 640 p.
13. Updike J. Rabbit, Run. – Greenwich, Conn.: Fawcett Publications, 1962. – 255 p.
14. Апдайк Дж. Кролик, беги // Апдайк Дж., Рот Ф. Романы / пер. М. Беккер. – М.: Физкультура и спорт, 1991. – С. 15-233.
15. The Moscow Tribune. – March 23, 1994.
16. International Herald Tribune. – March 23, 1994.
17. Shaw I. Rich Man, Poor Man. – Sevenoaks: New English Library, 1985. – 767 p.
18. Шоу И. Богач, бедняк // Шоу И. Богач, бедняк. Нищий, вор / пер. с англ. И. Басавиной. – М.: Радуга, 1986. – С. 19-440.
19. Street J. The Painless Method // Baker's Dozen: Thirteen Stories by Modern British and American Writers. – М.: Vysšaja Škola, 1979. – P. 47-51.
20. Стрит Дж. Пломба для мистера Смита / пер. с англ. М. Бирман // Неделя. – 1993. – № 46. – С. 14.
21. Trager G.L. Think Metric // On Language, Culture, and Religion: In Honor of Eugene A. Nida / eds. M. Black, W.A. Smalley. – The Hague-Paris: Mouton, 1974. – P. 373-380.
22. McCrum C. An English-Speaking World. – Harmondsworth: Penguin Books, 1993. – 342 p.
23. Хайруллин В.И. Корпусный словарь фобии-терминологии // Научно-техническая информация. Сер. 2. – 2013. – № 4. – С. 30-32.

Материал поступил в редакцию 08.12.14.

Сведения об авторе

ХАЙРУЛЛИН Владимир Иксанович – доктор филологических наук, профессор, директор центра повышения квалификации Института права Башкирского государственного университета, г. Уфа
E-mail: vladimir-blt@mail.ru

УВАЖАЕМЫЕ КОЛЛЕГИ!

ВИНИТИ РАН предлагает Вашему вниманию Реферативный Журнал в электронной форме

РЖ в электронной форме (ЭлРЖ) выпускается по всем разделам естественных, технических и точных наук.

Каждый номер ЭлРЖ является полным аналогом печатного номера РЖ по составу описаний документов, их оформлению и расположению. Он сопровождается оглавлением, указателями.

ЭлРЖ представляет собой информационную систему, снабженную поисковым аппаратом и позволяющую пользователю на персональном компьютере:

- читать номер РЖ, последовательно листая рефераты;
- просматривать рефераты отдельных разделов по оглавлению;
- обращаться к рефератам по указателям авторов, источников, ключевых слов;
- проводить поиск документов по словам и словосочетаниям;
- выводить текст описаний документов во внешний файл.

ЭлРЖ в версии Windows Вы можете получить за текущий год с любого номера, а также за предыдущие годы.

Подробную информацию Вы можете получить:

Адрес: 125190, Россия, Москва, ул. Усиевича, 20, ВИНТИ РАН

Телефон: 8 (499) 155-46-20

Телефон/Факс: 8 (499) 155-45-25

E-mail: zinovyeva@viniti.ru, davydova@viniti.ru

**Федеральное государственное бюджетное учреждение науки
ВСЕРОССИЙСКИЙ ИНСТИТУТ НАУЧНОЙ И ТЕХНИЧЕСКОЙ
ИНФОРМАЦИИ РОССИЙСКОЙ АКАДЕМИИ НАУК**

предлагает научным работникам, аспирантам и другим специалистам в области естественных, точных и технических наук, желающим быстро и эффективно опубликовать результаты своей научной и научно-производственной деятельности, использовать способ публикации своих работ через *систему депонирования*.

«Депонирование (передача на хранение) – особый метод публикации научных работ (отдельных статей, обзоров, монографий, сборников научных трудов, материалов научных конференций, симпозиумов, съездов, семинаров) узкоспециального профиля, разрешенных в установленном порядке к открытому опубликованию, широкое тиражирование которых, как правило, в силу их узкой специализации, не считается целесообразным, а также работ широкого профиля, срочная информация о которых необходима для утверждения их приоритета. Депонирование предусматривает прием, учет, регистрацию, хранение научных работ и обязательное размещение информации о них в специальных информационных изданиях».

Подготовка и передача на депонирование научных работ происходит в соответствии с «Инструкцией о порядке депонирования научных работ по естественным, техническим, социальным и гуманитарным наукам» (М., 2013).

Депонированные научные работы находятся на хранении в депозитарии ВИНТИ РАН, копии работ предоставляются заинтересованным организациям и специалистам на бумажном и электронном носителях и являются официальной публикацией.

Информация о депонированных научных работах включается в информационные издания ВИНТИ РАН, в РЖ ВИНТИ РАН и БД ВИНТИ РАН и аннотированный библиографический указатель «Депонированные научные работы».

Подать научную работу на депонирование можно, обратившись в Отдел депонирования ВИНТИ РАН по адресу:

125190, Москва, ул. Усиевича, 20.

ВИНТИ РАН, Отдел депонирования научных работ.

Тел.: 8 (499) 155-43-28, Факс: 8 (499) 943-00-60.

e-mail: dep@viniti.ru

С инструкцией о порядке депонирования можно ознакомиться на сайте ВИНТИ РАН: <http://www.viniti.ru>

Центр (Отдел) научно-информационного обслуживания (ЦНИО) ВИНИТИ РАН

предлагает услуги по предоставлению информационно-аналитических обзоров

ВИНИТИ РАН осуществляет подготовку информационно-аналитических обзоров по инновационным и приоритетным направлениям научных исследований в области точных, естественных и технических наук. Обзоры готовятся ведущими специалистами ВИНИТИ, работающими в определенных областях науки и техники. Аналитические материалы содержат результаты анализа и обобщения информации по актуальным научным проблемам, а в некоторых случаях – и прогностические выводы. Основой для составления обзоров служит отечественная и зарубежная научно-техническая литература, доступная ВИНИТИ РАН: фонд НТЛ, включающий более 2 млн отечественных и иностранных журналов, книг, депонированных рукописей, авторефератов диссертаций и другой научной литературы, ретроспектива – с 1987 года. Имеется доступ к базам данных и Интернет-ресурсам: БД ВИНИТИ (разработка ВИНИТИ), БД SCOPUS, БД зарубежных патентов и другим. Кроме того, ВИНИТИ доступны зарубежные электронные платформы ряда ведущих научных издательств, выпускающих основную часть академических рецензируемых журналов, в полнотекстовом варианте.

Основные тематические направления предлагаемых обзоров:

- Науки о жизни;
- Физико-математические науки;
- Химия и науки о материалах;
- Индустрия наносистем и материалов;
- Науки о Земле;
- Рациональное природопользование;
- Информационно-телекоммуникационные системы;
- Энергетика, энергоэффективность, энергосбережение;
- Транспортные, авиационные и космические системы;
- Производственные технологии.

Предлагается подготовка и заказ информационно-аналитических обзоров и материалов по тематике заказчика. Такие обзоры могут относиться к упомянутым выше тематическим направлениям, но могут иметь и междисциплинарный характер. В этом случае обзоры отражают актуальную научную информацию и научные достижения, происходящие на стыке наук.

Более подробная информация о приобретении, заказе и цене обзоров представлена на сайте ВИНИТИ www.viniti.ru

Приобретение и заказ обзоров от юридических лиц проводится на договорной основе. Форма договора для последующего оформления представлена на сайте ВИНИТИ.

Оформление договоров и других необходимых документов производится Центром научно-информационного обслуживания ВИНИТИ (ЦНИО). Возможен прием заказов от физических лиц, оплата производится на расчетный счет или в кассу ВИНИТИ РАН.

Выполненные в ВИНИТИ обзоры предоставляются заказчикам в печатном виде либо в электронном варианте после оплаты заказа.

Обращаться в ЦНИО ВИНИТИ:

- адрес: 125190, Россия, г. Москва, ул. Усиевича, 20.
- телефоны: 8(499) 155 -42 -43, 8(499) 155 -42 -17
- эл. почта cnio@viniti.ru, fdk@viniti.ru.
- факс 8(499) 930 -60 -00 (для ЦНИО).