Н. Г. Швец (Минск, МГЛУ)

ОСНОВНЫЕ ЭТАПЫ РАЗРАБОТКИ ФОРМАЛЬНОЙ МОДЕЛИ ВЗАИМОСВЯЗИ ВЕРБАЛЬНЫХ И НЕВЕРБАЛЬНЫХ СОСТАВЛЯЮЩИХ РЕКЛАМНОГО ОБЪЯВЛЕНИЯ

В статье рассматривается схема построения компьютерной модели, позволяющей к заданному тексту рекламного объявления подобрать наиболее подходящую по содержанию иллюстрацию. При этом основное содержание текста предлагается представлять в виде определенного набора главных и второстепенных ключевых слов, а содержание иллюстрации – в виде многоуровневого комплекса дескрипторов. Близость этих содержаний (и соответственно корреляцию между изображением и вербальным компонентом рекламного объявления) можно определить через максимальное число совпадений опорных слов и дескрипторов с учетом их степени важности.

Компьютерная реализация формальной модели показала ее достаточно высокую эффективность. Таким образом, было доказано, что формальная взаимосвязь вербального текста и иллюстрации рекламного объявления вполне осуществима на лексико-семантическом уровне.

В связи со стремительным развитием современных информационных технологий, направленных на адекватное решение проблем автоматизированной обработки семиотически неоднородных массивов информации, особый интерес в лингвистике представляет исследование механизмов порождения и восприятия смысла креолизованных текстов (в том числе и печатных рекламных объявлений), в которых в рамках единого сообщения сплетены вербальные и невербальные компоненты.

Конкретизируя в целях нашего исследования понятие *рекламное объявление* (РО), будем называть им семиотически неоднородный текст, содержащий вербальный (словесный) компонент (непосредственно рекламный текст) и визуальный (невербальный) компонент (изображение), представленный в письменной форме, заранее подготовленный, обладающий автономностью, направленный на донесение до адресата определенной информации с целью привлечения внимания к тому или иному виду товара.

Гипотеза исследования 1000 печатных РО по теме «Косметика и парфюмерия» состоит в том, что, представляя содержание текста РО в виде определенного набора главных и второстепенных опорных (или ключевых) слов, а содержание иллюстрации РО – в виде многоуровневого комплекса дескрипторов, можно определить близость этих содержаний через максимальное число совпадений опорных слов и дескрипторов с учетом их степени важности.

Общая схема построения компьютерной модели включает следующие основные этапы:

1) постановка задачи;

2) разработка модели;

3) проведение компьютерного эксперимента.

Под моделью в компьютерной лингвистике понимается формализованное описание ряда существенных лингвистических свойств объекта, системы нескольких объектов, процесса или явления, обладающее структурным или функциональным подобием [1, с. 94]. Такое описание может быть выражено конечным набором предложений какого-либо языка, математическими формулами, таблицами, графиками, специальными знаками или какими-нибудь схемами.

Рассмотрим подробнее перечисленные выше этапы разработки формальной модели.

Начнем с постановки задачи. Из 1000 проанализированных РО выбраны 200 РО, которые относятся к трем предметным областям: «Шампунь», «Крем для лица», «Краска для волос». Каждое РО состоит из двух основных частей: вербальной, содержащей текст рекламы, и изображения.

Необходимо для текста любого из упомянутых РО подобрать наиболее подходящее по содержанию изображение.

Каждая компьютерная модель опирается на некоторую базу данных (БД). В нашем исследовании она состоит из:

1) таблиц основного содержания (ТОС) исследуемых рекламных текстов;

2) формальных представлений изображений исследуемых РО, заданных в виде их тезаурусных описаний;

3) текстов РО;

4) изображений РО.

Списки опорных слов, тезаурусные описания, а также тексты и иллюстрации рекламных объявлений в БД делятся на 3 группы, соответствующие следующим:

1) шампунь;

2) крем для лица;

3) краска для волос.

В нашем исследовании при выделении ключевых слов текста учитывается абсолютная частота употребления знаменательных слов (с учетом всех их возможных синонимов и замен) и количество абзацев, в которых они встретились. В целях получения более качественного результата при выявлении основного содержания текстов исследуемых РО мы использовали статистический метод в сочетании с позиционным методом извлечения ключевых слов из текста. Преимущество выбранной нами методики состоит в возможности классифицировать слова конкретного текста в зависимости от степени их важности для семантической структуры текста по нескольким группам, а также в гибкой применимости данной методики к текстам РО с разным количеством абзацев.

Были получены ТОС для каждого исследуемого текста, которые дополнились ключевыми словами из заголовков (КСЗ) соответствующих РО. В ТОС опорные слова в соответствии с предметными свойствами своих референтов в общем случае могут быть разделены на следующие группы:

1) слова-объекты;

2) слова-признаки;

3) слова-действия;

5) прочие слова.

Например, в ТОС (табл. 1) текста РО № 1 (рис. 1) ГОС и КСЗ разделены на следующие группы:

1) слова- объекты: *шампунь*, *волосы*, *лаборатория*, *фрукты*, *концентрат*;

2) слова-признаки: *активный*, *укрепляющий*;

3) прочие слова: *сила*, *блеск*.

Таблица 1

Основное содержание текста РО № 1

Тип опорных слов	**Опорные слова текста**			
	объекты	признаки	действия	прочие
ГОС1	*шампунь*			
ГОС2	*волосы*			
ГОС3	*лаборатория*			
ГОС4	*фрукты*			
ГОС5	*концентрат*			
ГОС:		*активный*		
ГОС7				*сила*
ГОС8				*блеск*
КСЗ		*укрепляющий*		

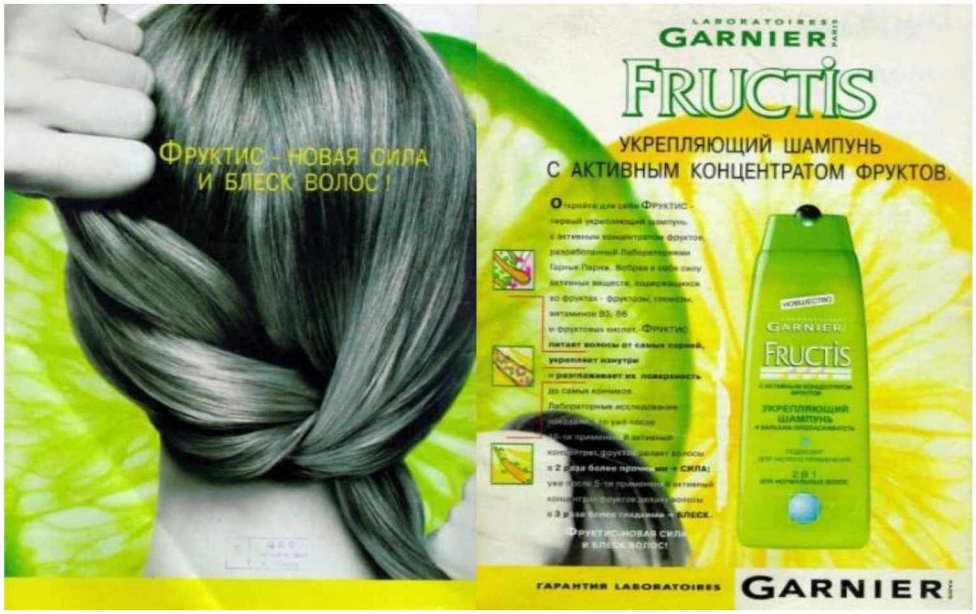


Рис. 1. Рекламное объявление № 1

Тезаурусное представление изображения РО № 1 приведено в табл. 2.

Таблица 2

Тезаурусное представление изображения РО № 1

Дескриптор (его уровень тезаурусной иерархии)	**Мероним (его уровень тезаурусной иерархии)**	**Признак (его уровень тезаурусной иерархии)**	**Действие (его уровень тезаурусной иерархии) (ассоциация)**
женщина (1)	*голова (2)* *руки (2)*		
флакон (1)	*шампунь (2)*	*зеленый (2)* *прямоугольный (2)*	
фон (1)	*фрукты (2)*		
схема-пояснение (1)	*волос (2)* *вещество (2)*		
голова (2)	*волосы (3)*		
руки (2)			*завязывать волосы в узел (3)* *(сила волос)*
шампунь (2)		*fructis (фруктис) (3)* *garnier paris (3)*	
волос (2)	*корень (3)*		
вещество (2)	*молекулы (3)*	*активное (3)* *зеленое (3)*	
фрукты (2)	*дольки (3)*		
волосы (3)		*блестящие (4)* *сильные (4)* *густые (4)* *шелковистые (4)* *длинные (4)* *гладкие (4)*	
корень (3)		*здоровый (4)* *сильный (4)*	
молекулы (3)		*зеленые (4)* *красные (4)*	*проникать в корень (4)* *(сила и блеск волос)*
дольки (3)		*зеленые (4)* *желтые (4)*	

Тезаурус является иерархической структурой, состоящей из дескрипторов различных уровней. Первый уровень тезауруса иллюстрации анализируемого РО (уровень тезауруса в табл. 2 обозначается цифрой в скобках) образуют следующие дескрипторы: *женщина*, *флакон*, *фон* и *схема-пояснение*.

Между дескриптором 1-го уровня *женщина* и дескрипторами 2-го уровня *голова* и *руки* устанавливаются системные отношения типа ЦЕЛОЕ (холоним) – ЧАСТЬ (мероним).

Дескриптор 1-го уровня *флакон* связан с дескриптором 2-го уровня *шампунь* отношением ЦЕЛОЕ – ЧАСТЬ, а с дескрипторами 2-го уровня *зеленый* и *прямоугольный* – отношением ОБЪЕКТ – ПРИЗНАК.

Между дескриптором 1-го уровня *схема-пояснение* и дескрипторами 2-го уровня *волос* и *вещество* устанавливаются отношения типа ЦЕЛОЕ – ЧАСТЬ.

Третий уровень тезауруса образуют следующие дескрипторы: *волосы*, *завязывать волосы в узел*, *сила волос*, *Fructis (Фруктис)*, *Garnier Paris*, *корень*, *молекулы*, *активное*, *зеленое*, *дольки*, *волосы*.

Дескрипторами 4-го уровня тезауруса являются: *блестящие*, *сильные*, *густые*, *шелковистые*, *длинные*, *гладкие*, *здоровый*, *сильный*, *зеленые*, *красные*, *желтые*, *проникать в корень*, *сила и блеск волос*.

Основным критерием максимальной близости по содержанию любого текста РО и иллюстрации, взятых из БД, является максимальное количественное совпадение ключевых слов текста с дескрипторами тезаурусного описания изображения, выбранными определенным способом.

Формальная модель процесса выбора иллюстрации к заданному тексту РО может быть представлена в виде принципиальной схемы алгоритма, основные блоки которого приведены на рис. 2.

Основной принцип работы алгоритма заключается в следующем.

Пользователь по своему желанию может выбрать на экране в меню предмет рекламы: «Крем для лица», «Шампунь» или «Краска для волос» (блок I), а затем – текст РО (блок II).

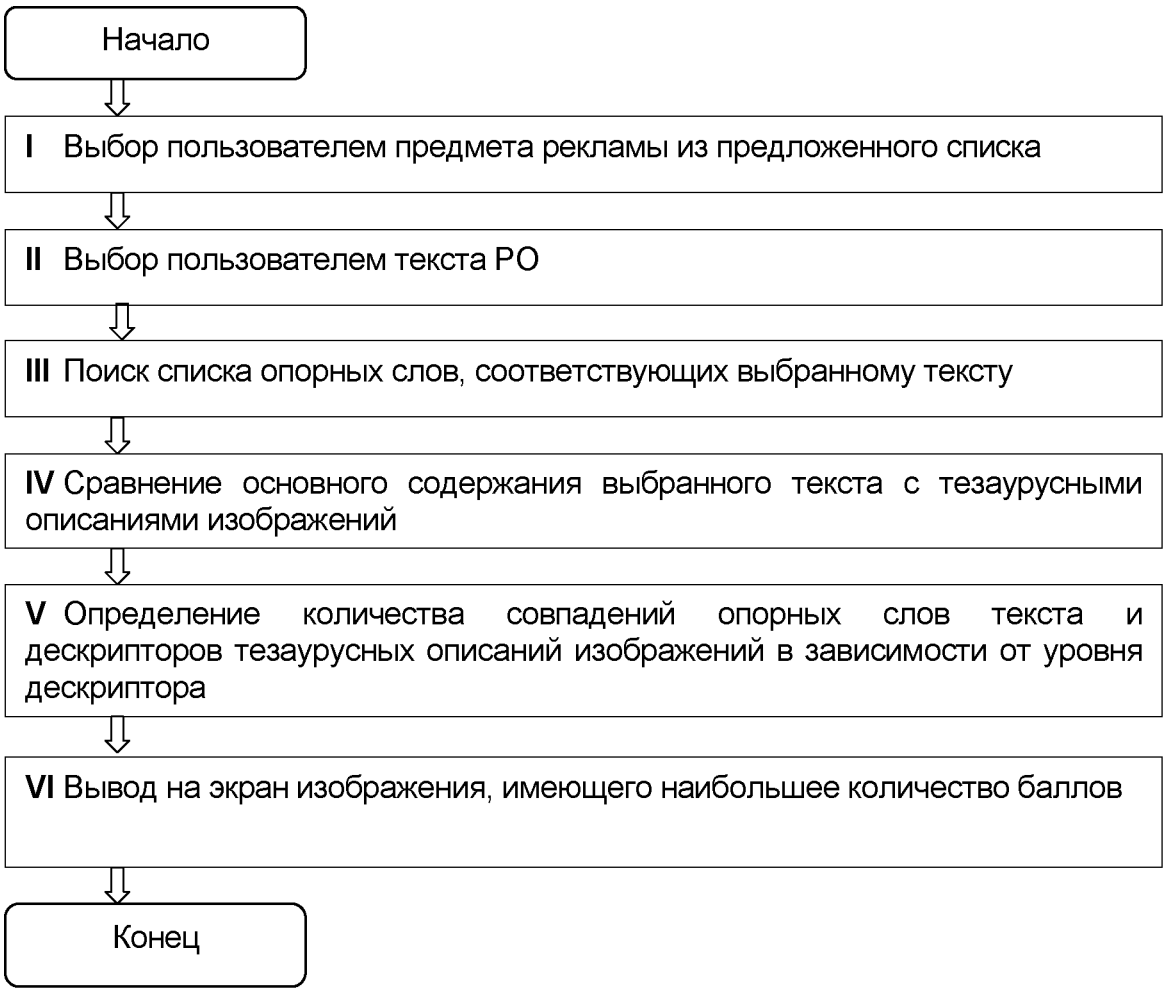


Рис. 2. Основные блоки принципиального алгоритма выбора иллюстрации к заданному тексту РО

Далее компьютер в автоматическом режиме подбирает иллюстрацию к данному тексту и высвечивает ее на экране. Для этого сначала осуществляются поиск в БД списка опорных слов, соответствующих выбранному тексту

(блок III), и его сравнение с тезаурусным описанием каждой иллюстрации, относящейся к выбранному предмету рекламы (блок IV). Затем определяется количество совпадений опорных слов текста и дескрипторов всех иллюстраций (блок V). В результате на экран выводится иллюстрация, которая имеет наибольшее количество баллов, которое зависит не только от количества совпадений, но и от уровня дескриптора (блок VI).

Для проведения компьютерного эксперимента была написана программа на языке C#.

Программа работает следующим образом. Сначала из меню пользователь выбирает один из предметов рекламы (рис. 3), затем – один из текстов, описывающих выбранный предмет рекламы.

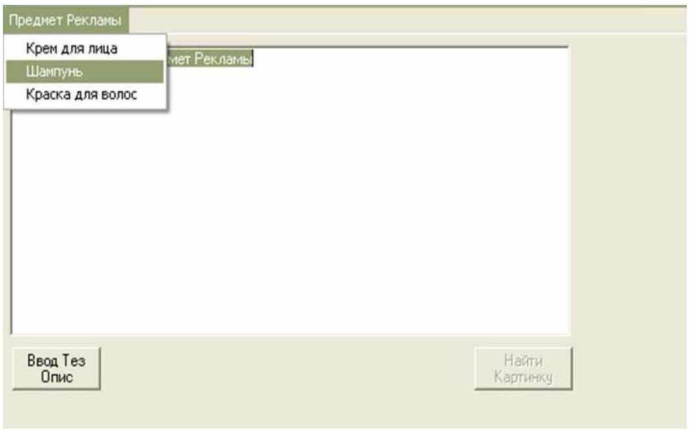


Рис. 3. Выбор из меню предмета рекламы

После нажатия на кнопку «Найти картинку» (рис. 4) на экране высвечивается таблица с результатами совпадения опорных слов и дескрипторов всех иллюстраций БД. В этой таблице перечисляются опорные слова выбранного текста РО, а также совпавшие с ними дескрипторы определенного уровня из тезаурусных описаний иллюстраций. Каждому такому совпадению начисляется определенное количество баллов, которое зависит от уровня дескриптора.

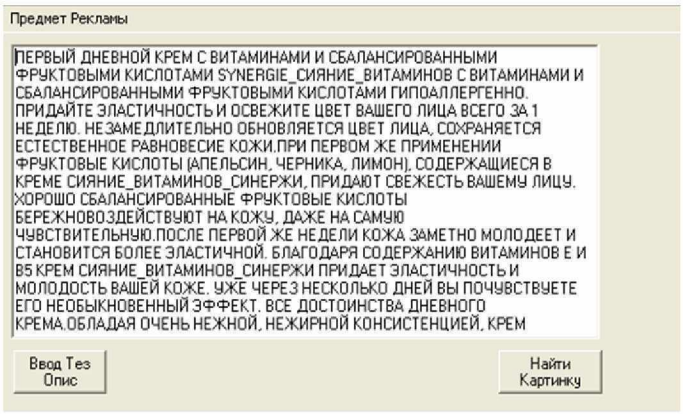
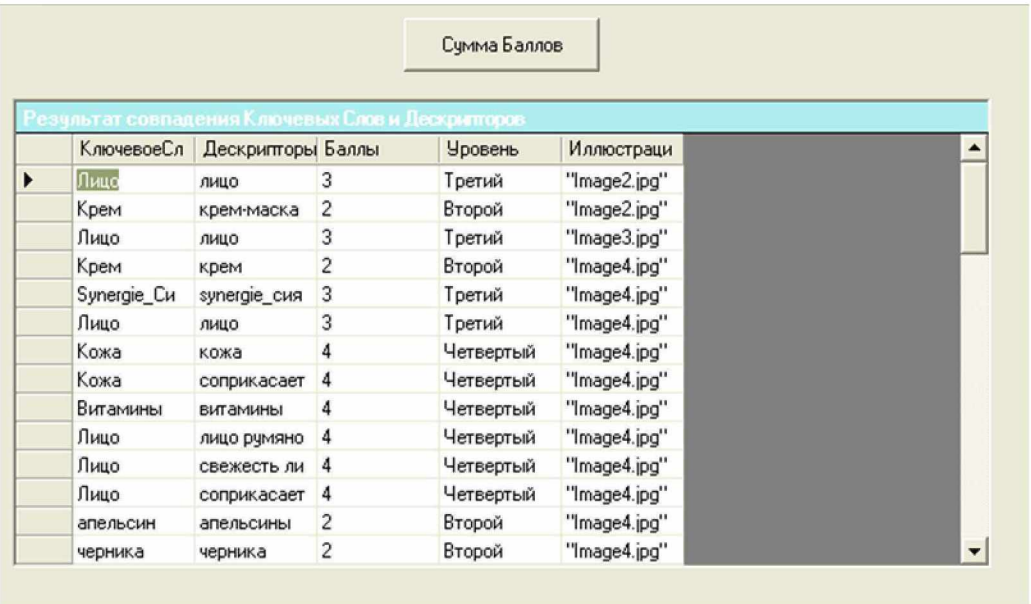


Рис. 4. Окно с выбранным текстом

Далее при нажатии на кнопку «Сумма баллов» (рис. 5) подсчитывается количество совпадений опорных слов текста и дескрипторов, описывающих изображения РО, и на экран выдается иллюстрация, имеющая наибольшее значение числа таких совпадений.



КлючевоеСл	Дескрипторы	Баллы	Уровень	Иллюстраци
Лицо	лицо	3	Третий	"Image2.jpg"
Крем	крем-маска	2	Второй	"Image2.jpg"
Лицо	лицо	3	Третий	"Image3.jpg"
Крем	крем	2	Второй	"Image4.jpg"
Synergie_Си	synergie_сия	3	Третий	"Image4.jpg"
Лицо	лицо	3	Третий	"Image4.jpg"
Кожа	кожа	4	Четвертый	"Image4.jpg"
Кожа	соприкасает	4	Четвертый	"Image4.jpg"
Витамины	витамины	4	Четвертый	"Image4.jpg"
Лицо	лицо румяно	4	Четвертый	"Image4.jpg"
Лицо	свежесть ли	4	Четвертый	"Image4.jpg"
Лицо	соприкасает	4	Четвертый	"Image4.jpg"
апельсин	апельсины	2	Второй	"Image4.jpg"
черника	черника	2	Второй	"Image4.jpg"

Рис. 5. Окно с результатами совпадения опорных слов текста и дескрипторов тезаурусных описаний иллюстраций

Анализ результатов работы компьютерной программы показал ее достаточно высокую эффективность. Таким образом, было доказано, что формальная взаимосвязь вербального текста и иллюстрации РО вполне осуществима на лексико-семантическом уровне.

Предложенные идеи можно использовать при создании информационных систем семантического поиска визуальных материалов в криминалистике (при отождествлении предметов и их описаний) и в музейном деле (при составлении документации на картины и другие экспонаты), в издательском деле (при компьютерном дизайне текстов), а также в вузовских учебных курсах по лингвистике текста, семиотике и прикладной лингвистике.

ЛИТЕРАТУРА

1. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Е. И. Большакова [и др.]. – М. : МИЭМ, 2011. – 272 с.

The article discusses the scheme for constructing a computer model that allows to select the most suitable illustration for the content of a given ad text. The computer implementation of the formal model showed its rather high efficiency. Thus, it was proved that the formal interrelation of the verbal text and the illustration of an advertisement is quite feasible at the lexical-semantic level.