УДК 811.111’33(045)

Н. А. МЕТЛИЦКАЯ

ЛИНГВИСТИЧЕСКАЯ БАЗА ДАННЫХ СИСТЕМЫ АВТОМАТИЧЕСКОГО ПОРОЖДЕНИЯ АНГЛОЯЗЫЧНОГО РЕКЛАМНОГО ТЕКСТА

Минский государственный лингвистический университет

Целью данной работы является разработка лингвистического обеспечения автоматической системы порождения англоязычного рекламного текста по косметике и парфюмерии, и ее последующая реализация в виде компьютерной программы. Создаваемая система разрабатывается по принципу лингвистически мотивированных технологий, что требует использования широкого спектра лингвистических знаний о структуре и содержании порождаемого текста (базы данных, семантические и формальные языки). Лингвистическая база данных рассматриваемой системы включает следующие компоненты: автоматический словарь лексических единиц с указанием семантических и морфологических сведений, семантико-синтаксические формулы текстов на формальном языке СЕМСИНТ. В работе рассматривается каждая составляющая этой базы данных.

Словарь лексических единиц строится на основе анализа тридцати оригинальных англоязычных рекламных текстов по косметике и парфюмерии, относящимся к трем предметным областям (губная помада, тушь для ресниц, шампунь). Словарная статья автоматического словаря включает две зоны: зону грамматических сведений, зону семантических сведений. Зона грамматических сведений содержит информацию о части речи лексической единицы, а также набор ее морфологических признаков. Зона семантических сведений включает семантический признак лексической единицы, т. е. ее отнесенность к определенному семантическому подклассу. Для этого была произведена семантическая классификация всех слов исследуемых текстов с присвоением им соответствующих кодов. В качестве примера в работе приводится результат семантической классификации имен существительных исследуемых рекламных текстов предметной области «губная помада».

Вторую часть базы данных составляют семантико-синтаксические формулы текстов на формальном языке СЕМСИНТ. В работе описываются составляющие языка СЕМСИНТ, а также рассматривается его сущность и правила его использования. Представлен пример семантико-синтаксической формулы текста, созданной средствами данного формального языка.

Ключевые слова: *автоматическое порождение рекламного текста, лингвистическое обеспечение, автоматический словарь лексических единиц, формальный язык.*

Введение

Несмотря на то, что автоматическим порождением текста занимаются уже на протяжении нескольких десятилетий, данная проблема все еще находится на стадии исследования и разработки [1, 2, 3, 4 и др.]. Связано это с тем, что текст представляет собой сложное, многоуровневое явление, результат взаимодействия большого числа факторов – лингвистических, психологических, логических, коммуникативных. На основе результатов исследований в общем виде можно выделить два принципиально различных подхода к порождению текста компьютером [5, 6]:

– системы автоматического порождения текста, основанные на использовании шаблонов, т. е. шаблонные технологии;

– системы автоматического порождения текста, основанные на использовании лингвистических знаний, т. е. лингвистически мотивированные технологии.

Первые предполагают использование при порождении текста готовых текстовых фрагментов, и потому имеют ряд ограничений, связанных с невозможностью менять формат и содержание порождаемого текста. Вторые работают с содержанием будущего текста, представленном в виде данных нетекстовой природы (базы данных, базы знаний, семантические и формальные языки) [6]. Они подраз-

умевают моделирование процесса порождения текста человеком, что требует использования широкого спектра лингвистических знаний. Данная работа посвящена разработке лингвистического обеспечения автоматической системы порождения англоязычного рекламного текста по косметике и парфюмерии.

База данных системы автоматического порождения англоязычного рекламного текста

Лингвистическая база данных системы автоматического порождения рекламного текста включает следующие компоненты: 1) автоматический словарь лексических единиц с указанием семантических и морфологических сведений; 2) семантико-синтаксические формулы текстов на формальном языке СЕМСИНТ.

Рассмотрим каждую составляющую описываемой базы данных подробнее.

Словарь лексических единиц разрабатываемой системы строится на основе анализа тридцати оригинальных англоязычных рекламных текстов по косметике и парфюмерии, относящимся к трем предметным областям (губная помада, тушь для ресниц, шампунь). Тексты взяты с официальных сайтов косметических компаний сети Интернет. Словарная статья автоматического словаря включает две зоны:

- семантических сведений;
- грамматических сведений.

Зона семантических сведений содержит семантический признак лексической единицы, т. е. ее отнесенность к определенному семантическому подклассу. Разработанная в данном исследовании семантическая классификация лексических единиц основана на денотативно-сигнификативном принципе и отражает классификацию самих предметов и явлений, обозначенных данными словами. Семантические подклассы выделены в составе классов лексических единиц, имеющих общую категориальную сему предметности (имена существительные), признаковости (имена прилагательные, наречия), процессуальности (глаголы). В табл. 1 представлена общая схема семантической классификации

Таблица 1. **Семантическая классификация имен существительных рекламных текстов по косметике и парфюмерии (губная помада)**

Сегменты реальной действительности	Семантический подкласс существительного	Код подкласса
Человек и его мир	Понятия, обозначающие части тела человека (lip, pout)	N01
	Понятия, описывающие части тела человека (surface, finish)	N02
	Понятия, используемые для обозначения действия (application, coverage, wear)	N03, N04
	Понятия, связанные с состояниями, ощущениями, чувствами восприятия (feel, look)	N05, N06
	Понятия, используемые для положительной характеристики человека и неживой природы (beauty, sophistication, care, caress, indulgence)	N07
Природа	Состояния и процессы, протекающие в природе, имеющие положительный результат (moisture, comfort)	N08
	Состояния и процессы, протекающие в природе, имеющие негативный результат (fading, dulling, feathering, drying)	N09
	Вещества, материалы, природные элементы (Vitamin, oil, wax, pigment, Omega)	N10, N11, N12, N13
	Растения (argan)	N14
	Цвета, оттенки (color, shade, rose, red, coral, scarlet)	N15, N16, N17
	Абстрактные понятия, мыслимые образы (experience, effect, balance)	N18, N19
Производство	Косметические средства (lipstick, lipcolour, gel-color)	N20
	Понятия, характеризующие косметические средства по составу/структуре (system, formula, texture)	N21
Предметы	Предметы-артефакты и их части (applicator, tube)	N22
	Физические характеристики предметов (gloss, shine, luster, shimmer)	N23
	Качественные характеристики предметов (smoothness, creaminess)	N24
Время	Временные понятия (hour, day)	N25, N26
Имена собственные	Наименования (E)	N27

имен существительных исследуемых рекламных текстов, относящихся к предметной области «губная помада». С более подробной лексико-семантической классификацией исследованных текстов можно ознакомиться в работе [7].

Зона грамматических сведений содержит информацию о части речи лексической единицы, а также набор ее морфологических признаков. Для этого каждой лексической единице присваивается определенный код, обозначающий ее часть речи. Так, например, кодом N обозначаются имена существительные, A – прилагательные, V – глаголы, S – наречия, P – местоимения, M – числительные, и т. д.

Каждый класс знаменательных слов, составляющих словарь системы порождения, обладает собственными грамматическими признаками. Поэтому морфологическая информация, представленная в зоне грамматических сведений лексических единиц разных частей речи, различна. Так, например, в словарной статье имени существительного приводятся формы единственного и множественного числа. В зоне морфологических сведений для глагола указываются формы инфинитива, 3-го лица единственного числа настоящего времени, формы причастий. Указания более подробной информации в силу специфики употребления в английском языке требует глагол *to be*. Примеры словарных статей упомянутых частей речи представлены в табл. 2, 3, 4.

Таблица 2. **Схема словарной статьи имени существительного**

Основа	Форма множественного числа	Код семантического подкласса
1	2	3
color	colors	N15

Таблица 3. **Схема словарной статьи глагола**

Инфинитив	Форма 3-го лица ед. ч. настоящего времени	Причастие I	Причастие II	Код семантического подкласса
1	2	3	4	5
discover	discovers	discovering	discovered	V01

Таблица 4. **Схема словарной статьи глагола *to be***

Инфинитив	Формы наст. времени			Причастие I	Причастие II	Код семантического подкласса
1	2			3	4	5
	1	2	3			
be	am	is	are	being	been	V18

Одной из задач, которую необходимо решить при разработке системы автоматического порождения текста, является представление содержания будущего текста на некотором формальном языке, понятном компьютеру. Семантико-синтаксический язык СЕМСИНТ позволяет формально описывать содержание текста с учётом семантико-синтаксических отношений и связей между его компонентами. Набор семантико-синтаксических формул исследуемых текстов на языке СЕМСИНТ представляет вторую составляющую базы данных рассматриваемой системы порождения текста.

1) В самом общем виде язык СЕМСИНТ включает следующие компоненты:

2) алфавит языка (все знаки основного набора клавиатуры современных персональных компьютеров, т. е. русские и латинские буквы, цифры, орфографические знаки, знаки арифметических и логических действий);

3) систему средств для записи семантических отношений между членами предложения;

4) систему средств для записи синтаксических отношений между членами предложения.

Для описания смысла предложения используется падежно-ролевой подход. Автор [8] исходит из идеи о том, что предложение (текст) отражает некоторую ситуацию действительности, а семантическая формула должна отражать семантико-синтаксические роли конкретных участников ситуации, описанной в тексте. В основе данного формального языка лежит идея о глубинных падежах Чарльза Филлмора. Как известно, суть идеи Ч. Филлмора заключается в том, что семантические падежи отображают глубинную семантическую структуру предложения, указывая на роли конкретных участников реальной ситуации или события, которые описываются в предложении (тексте) [9]. В работе [9] предлагается использовать семь глубинных падежей:

– агентив (действующее лицо, производитель действия);

– датив (адресат);

– инструменталис (инструмент);

– фактитив (неодушевленный предмет, возникающий в результате действия, означаемого глаголом);

– локатив (место);

– объектив (неодушевленный объект, подвергающийся воздействию);

- темпоратив (время).

В формальном языке СЕМСИНТ для фиксации семантических отношений между компонентами текста используется определенный набор семантических функций, подобных семантическим падежам Ч. Филлмора. Нами были использованы 19 семантических функций, описанных в работе [8]: AAG – субъект, AH1 – предмет, AEL – П-деятель, AP – объект, AB2 – адресат, AO – Н-объект, ALK – место, AMD – средство, AIN – инструмент, AKM – состав, AD1 – определитель, ATM – время, AAD – способ, CND – условие, PPS – цель, RST – результат, CAS – причина, PRP – свойство, SPC – уточнение. Данные функции выделены с учетом специфики исследуемых текстов.

Следует отметить, что семантические функции относятся не к отдельному слову предложения, а к аргументной группе – группе слов, включающих в себя существительное и все относящиеся к нему определители, выраженные прилагательными, причастиями, числительными, местоимениями. В семантико-синтаксической формуле предложения аргументная группа представлена цепочкой кодов семантических подклассов существительного и определителей, которые соединяются знаком «*». Слева и справа аргументная группа ограничивается знаками «<» и «>», соответственно. Перед каждой цепочкой ставится один из кодов семантических падежей. Для обозначения в семантической формуле глагола-сказуемого используется код *Rn*, где *n* – цифра, указывающая семантическую валентность глагола. Затем между знаками «<» и «>» указывается код семантического подкласса глагола. Все составные части семантической формулы предложения соединяются между собой знаком «+».

Создание семантико-синтаксических формул исследуемых текстов происходит на основе рассмотренного выше словаря лексических единиц и присвоенных им соответствующих кодов. Например, запись N01/2 означает, что данная лексическая единица является именем существительным, которое относится к семантическому подклассу «части тела человека» и употреблено в форме множественного числа. Для отображения синтаксических отношений между членами предложения соответствующие коды были присвоены предлогам и союзам.

Ниже приводится пример одного из исследованных рекламных текстов:

Color Rich Gloss

Extreme shine.

Supreme creaminess.

Ultra-hydrating argan oil & Omega 3.

Discover our rich shine indulgence for your lips.

New C*olor Rich Gloss is infused with ultra-hydrating Omega 3, protective vitamin E, and argan oil for our rich shine experience. Wear alone or over your favorite shade of lipcolour to customize your look. Available in 16 sensual shades.*

Семантико-синтаксическая формула данного текста на формальном языке СЕМСИНТ была представлена следующим образом:

AH1<N15/1*A12*N23/1>

AD1<A12*N23/1>.

AD1< A12*N24/1>.

AKM<A11*N14/1*N11/1*C01*N13/1*3>.

R<V01/1>+AO<P09*A12*N23/1*N07/1>+AD1<J05*P10*N01/2>.

AH1<A05*N15/1*A12*N23/1>+R2<V18/2.2*V06/4>+AKM<J03*A11* N13/1*3,*A11*N10/1*N27/1*C01*N14/1*N11/1> + PPS<J05* P09*A12*N23/1*N18/1>.

R<V02/1> + AAD<A18*C03*J08*P10*A14*N16/1*J04*N20/1>+R<K01*V03/1>+AO<P10*N06/1>. AD1<A13*J06*16*A05*N16/2>.

Заключение

Важным этапом создания системы автоматического порождения текстов является разработка лингвистических ресурсов, которые описывают знания, необходимые для текстопорождения. В работе представлены результаты разработки лингвистического обеспечения для системы автоматического порождения англоязычных рекламных текстов по косметике и парфюмерии. Алгоритм работы компьютерной программы порождения рекламного текста основан на взаимодействии описанных блоков базы данных. Использование представленной лингвистической базы данных позволит автоматически создавать англоязычные рекламные тексты, максимально приближенные по своему содержанию и семантико-синтаксической структуре к текстам, порождаемым человеком.

Литература

1. **Болдасов, М. В.** Генерация текстов на естественном языке – теории, методы, технологии / М. В. Болдасов, Е. Г. Соколова // НТИ. Сер.2, Информац. процессы и системы. – 2006. – № 7. – С. 1–14.

2. **Всеволодова, А. В**. Компьютерная обработка лингвистических данных: учеб. пособие / А. В. Всеволодова. – 2-е изд. – М.: Флинта: Наука, 2007. – 96 с.

3. **Пиотровский, Р. Г.** Текст, машина, человек / Р. Г. Пиотровский. – Л.: Наука, 1975. – 327 с.

4. **Mann, W. C.** Computer generation of multiparagraph English text / W. C. Mann, J. A. Moore // American J. of Computational Linguistics. – 1981. – Vol. 7 (1). – pp. 17–29.

5. **Бусел, Т. В.** Лингвистические аспекты автоматического порождения англоязычных деловых электронных писем: дис. …канд. филол. наук: 10.02.21 / Т. В. Бусел. – Минск, 2011. – 113 л.

6. **Соколова, Е. Г.** Автоматическая генерация текстов на ЕЯ / Е. Г. Соколова [электронный ресурс]. – Режим доступа: http:// www. dialog-21. ru/archive/2004/sokolova. htm. – Дата доступа: 08.04.2015.

7. **Метлицкая, Н. А.** Лексико-семантическая специфика англоязычных рекламных текстов по косметике и парфюмерии / Н. А. Метлицкая // Вестник МГЛУ. Сер. 1, Филология. – 2016. – № 3(82). – С. 109–116.

8. **Зубов, А. В.** Семантико-синтаксический язык для записи текстов в памяти ЭВМ / А. В. Зубов // Функционирование и развитие языковых систем. Сб. научных трудов. – Минск: Вышэйшая школа, 1990. – С. 110–117.

9. **Филлмор, Ч**. Дело о падеже / Ч. Филлмор // Лингвистика XX века: система и структура языка: хрестоматия в 2 ч., сост. Е. А. Красина. – М.: РУДН, 2004. – Ч. 2. – С. 75–96.

References

1. **Boldasov, M. V.** Text generation in natural language – theories, methods, technologies / M. V. Boldasov, E. G. Sokolova // NTI. Ser. 2, Inform. Processes and systems. – 2006. – № 7. – pp. 1–14.

2. **Vsevolodova, A. V.** Computer processing of linguistic data: Handbook / A. V. Vsevolodova. – 2nd ed. – Moscow: Flinta: Science, 2007. – 96 p.

3. **Piotrovsky, R. G.** Text, machine, human being / R. G. Piotrovsky. – L.: Science, 1975. – 327 p.

4. **Mann, W. C.** Computer generation of multiparagraph English text / W. C. Mann, J. A. Moore // American J. of Computational Linguistics. – 1981. – Vol. 7 (1). – pp. 17–29.

5. **Busel, T. V.** Linguistic aspects of automatic generation of English business correspondence texts: dissertation¼PhD in Philology: 10.02.21 / T. V. Busel. – Minsk, 2011. – 113 p.

6. **Sokolova, E. G.** Automatic text generation in natural language [Electronic resource]. – Mode of access: http://www.dialog-21.ru/archive/2004/sokolova.htm. – Date of access: 08.04.2015.

7. **Metlitskaya, N. A.** Lexical-semantic specificity of English advertising texts on cosmetics and perfumery / N. A. Metlitskaya // Vestnik MSLU, Vol. 3(82). – 2016. – pp. 109–116.

8. **Zubov, A. V.** A semantic and syntactic language for text entry in computer memory / A. V. Zubov // Functioning and development of language systems: Collection of scientific papers. – Minsk: Vyshejshaya shkola, 1990. – pp. 110–117

9. **Fillmore, Ch**. The case for case / Ch. Fillmore // Linguistics of XX century: language system and structure: hrestomatija in 2 parts, sost. E. A. Krasina. – Moscow: RUDN, 2004. – Part 2. – pp. 75–96.

Поступила	*После доработки*	*Принята к печати*
28.03.2017	*05.04.2017*	*10.06.2017*

N. A. Metlitskaya

LINGUISTIC DATABASE FOR AUTOMATIC GENERATION SYSTEM OF ENGLISH ADVERTISING TEXTS

Minsk State Linguistic University

The article deals with the linguistic database for the system of automatic generation of English advertising texts on cosmetics and perfumery. The database for such a system includes two main blocks: automatic dictionary (that contains semantic and morphological information for each word), and semantic-syntactical formulas of the texts in a special formal language SEMSINT. The database is built on the result of the analysis of 30 English advertising texts on cosmetics and perfumery. First, each word was given a unique code. For example, N stands for nouns, A – for adjectives, V – for verbs, etc. Then all the lexicon of the analyzed texts was distributed into different semantic categories. According to this semantic classification each word was given a special semantic code. For example, the record N01 that is attributed to the word «lip» in the dictionary means that this word refers to nouns of the semantic category «part of a human's body».

The second block of the database includes the semantic-syntactical formulas of the analyzed advertising texts written in a special formal language SEMSINT. The author gives a brief description of this language, presenting its essence and structure. Also, an example of one formalized advertising text in SEMSINT is provided.

Keywords: *automatic generation of English advertising texts, linguistic database, automatic dictionary, semantic-syntactical formal language.*

Метлицкая Наталья Анатольевна, аспирант кафедры информатики и прикладной лингвистики Минского государственного лингвистического университета (специальность 10.02.21 – прикладная и математическая лингвистика). Окончила Минский государственный лингвистический университет (2007), магистратуру МГЛУ (2008). С 2007 по 2013гг. работала преподавателем английского языка на кафедре интенсивного обучения иностранным языкам факультета иностранных языков для руководящих работников и специалистов МГЛУ. Научные интересы: прикладная лингвистика, лингвистическое обеспечение информационных систем.

Адрес: г. Минск, ул. Есенина, 6/3, кв.81. Тел: (29) 502 88 91.

E-mail: magnatusik@gmail.com

Natalia Metlitskaya, postgraduate student of Minsk State Linguistic University, department of Informatics and Applied linguistics (specialization 10.02.21 – Applied and mathematical linguistics). Graduated from Minsk State Linguistic University (2007), Master of Philology (2008). From 2007 to 2013 worked as an English teacher at the Department of Intensive learning of foreign languages of MSLU. Scientific interests: Applied linguistics, lingware for automatic systems. E-mail: magnatusik@gmail.com tel.: (29) 502 88 91.