

С. В. ПЕТРАСОВА, Н. Ф. ХАЙРОВА

## ЛОГИКО-ЛИНГВИСТИЧЕСКАЯ МОДЕЛЬ ИДЕНТИФИКАЦИИ СЕМАНТИЧЕСКИ ЭКВИВАЛЕНТНЫХ КОЛЛОКАЦИЙ

Предлагается логико-лингвистическая модель идентификации семантически близких коллокаций. Для автоматического определения семантической корреляции эквивалентности между коллокатами использован метод компонентного анализа. В предлагаемой модели для формализации семантически близких коллокаций определено множество семантико-грамматических характеристик коллокатов, отношения между которыми описаны с использованием базового аппарата алгебры предикатов. Сделаны выводы о необходимости учета как синтаксической, так и семантической информации о лексических единицах, что позволит повысить эффективность семантической обработки текстов.

**Ключевые слова:** семантически близкие коллокации, корреляция эквивалентности, коллокаты, метод компонентного анализа, семантико-грамматические характеристики, аппарат алгебры предикатов.

**Введение.** Семантический анализ текста – одно из самых динамично развивающихся и в то же время наиболее сложно реализуемых направлений в сфере Natural language processing (NLP). Сложность семантической обработки естественно-языковых источников, в первую очередь, определяется многозначностью и синонимичностью, присущими языку на всех уровнях его представления (морфологическом, синтаксическом, семантическом и прагматическом), что, прежде всего, проявляется в проблеме определения семантической эквивалентности языковых единиц. При этом сложность задачи повышается экспоненциально, если рассматривается смысловая близость не слов, а многословных словосочетаний.

Для установления смысловой эквивалентности словосочетаний предлагается логико-лингвистическая модель определения семантической близости. Применение данной модели позволяет автоматизировать определение семантической близости двухсловных коллокаций. Для идентификации коллокаций предлагается учитывать как морфологические и синтаксические, так и семантические характеристики исследуемых лексических единиц.

Под коллокацией в данном исследовании мы определяем комбинацию двух лексических единиц, имеющих тенденцию к совместной не случайной встречаемости в тексте.

**Актуальность исследования.** Семантические эквиваленты можно определить как слова с близким значением, встречающиеся в одном контексте (т.е. слова с совпадающим денотатом и различными сигнификатами). Таким образом, с помощью семантических эквивалентов осуществляется кодирование одного и того же содержания разными формальными средствами.

Понятие семантической эквивалентности (близости) неразрывно связано с контекстом. Так при определении близких по смыслу отдельных лексем (семантических эквивалентов) необходимо учитывать принадлежность их значений к одной и той же предметной или понятийной области. Тогда как, при определении семантической близости коллокаций или словосочетаний необходимо устанавливать семантическую близость конкретных смыслов коллокаций полностью, а не лексем, из которых состоят слово-

сочетания, со всей их индивидуальной многозначностью.

Существующие методы определения коллокаций можно разделить на две группы. К первой группе методов относятся статистические методы (window-based, меры ассоциации MI, PMI, t-scores, Chi-squared распределение), основанные на измерении частоты совместной встречаемости коллокатов. Вторая группа – методы, основанные на анализе синтаксической структуры коллокаций, в результате которого генерируется список коллокатов.

Методы "window-based" основаны на модели линейного порядка слов, в которых кандидаты коллокации (коллокаты) извлекаются из "окна" фиксированного размера [1]. Меры MI (Mutual information) и PMI (Pointwise mutual information) позволяют определить, насколько значимой является встречаемость двух слов на основе сравнения частоты совместного появления двух слов и произведением частот их независимого появления в тексте [2]. Мера t-score учитывает частоту совместной встречаемости ключевого слова и его коллоката. Слова с наибольшим значением t-score оказываются частотными, поэтому необходимо задавать список стоп-слов, чтобы отбросить самые частотные слова. В Chi-squared распределении используется  $\chi^2$ -критерий Пирсона для анализа таблиц сопряженности. Значения, формирующие таблицу:

- частота данной коллокации,
- частота коллокаций с участием первого слова (но не второго),
- частота коллокаций с участием второго слова (но не первого),
- частота всех остальных коллокаций [3].

Статистические методы, как правило, извлекают дополнительные шумные данные и игнорируют синтаксические связи между словами на длинных расстояниях.

Использование методов, основанных на анализе синтаксической структуры, позволяет отфильтровать ложные коллокаты, а также получить доступ к коллокатам, находящимся на длинном расстоянии друг от друга. Следует отметить, что такое увеличение точности достигается за счет тщательного описания всех синтаксических конструкций, в которых могут возникнуть два коллоката [4].

К настоящему моменту ни первая, ни вторая группа методов не позволяют выявлять эквивалентность словосочетаний с достаточно высокой полнотой. Анализ показывает, что определение коллокаций в текстах требует, в дополнение к семантико-синтаксическим средствам анализа, использования моделей когнитивной деятельности человека. В этом случае идентификация семантически близких коллокаций позволяет не только учитывать многозначность единиц языка, но и формализовать семантические корреляции, в частности, семантическую эквивалентность.

**Постановка задачи исследования.** Языковая деятельность человека характеризуется двумя основными классами когнитивных средств: языковыми и прагматическими. Первый класс – это знания, закрепленные в семантике языка; второй обеспечивается использованием неязыковых знаний в сочетании с языковыми [5], наличием пресуппозиции – необходимого семантического компонента, обеспечивающего существование смысла в утверждении. В связи с чем ни одна из задач, связанных с автоматической обработкой текста, не может быть решена исключительно на морфологическом и/или синтаксическом анализе, а требует более сложного (семантического) уровня обработки.

Цель данной работы – разработка логико-лингвистической модели идентификации семантически близких коллокаций на основе использования методов компонентного анализа, с использованием базового аппарата алгебры конечных предикатов.

Основная идея данного подхода заключается в том, что обладающие семантическими корреляциями коллокаты имеют определенную общность содержания, выражающую сходство обозначаемых явлений или понятий [6].

**Пример выявления семантически эквивалентных коллокаций.** На первом этапе семантической обработки текста должны быть выделены лексические семантические эквиваленты. Для этого используется метод компонентного анализа, основанный на использовании знаний глоссария [7].

Двухсловные словосочетания (коллокации), образованные попарно семантически близкими коллокатами, могут быть как семантически близкими, так и семантически не близкими, имеющими очень маленькую смысловую общность. Примеры семантически близких словосочетаний показаны на рис. 1, это пары – *контент ресурсу*  $\approx$  *зміст джерела*; *об'єднання споживачів*  $\approx$  *асоціація користувачів*; *часовий інтервал*  $\approx$  *проміжок часу*; *зберігати дані*  $\approx$  *тримати позначки*.

К семантически не близким словосочетаниям, составленным из семантически близких слов можно отнести: *виділятися форматом*  $\neq$  *позначати розмір*; *гуртування стандартами*  $\neq$  *поєднання шаблонів*; *оснащений кабелем*  $\neq$  *забезпечення дроту* (рис. 2).

Предлагается логико-лингвистическая модель, позволяющая формально определить семантическую эквивалентность двухсловных словосочетаний за счет

отношений семантико-грамматических признаков главных и зависимых коллокатов в субстантивной, адъективной и глагольной коллокациях.



Рис. 1 – Семантически эквивалентные коллокации

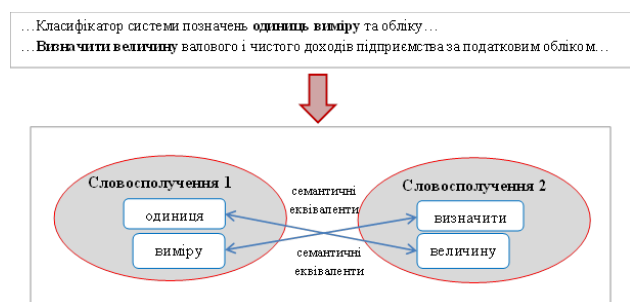


Рис. 2 – Семантически не эквивалентные коллокации

**Описание модели идентификации семантической эквивалентности коллокаций.** Рассмотрим множество словоформ, образующих словосочетание  $M = \{m_1, \dots, m_n\}$ , где  $n$  – количество словоформ в слове системы. Образующая коллокации, словоформы из множества  $M$  устанавливают семантико-синтаксические связи, которые можно выразить формально, используя базовые средства алгебры конечных предикатов, позволяющие идентифицировать лингвистические объекты.

Семантико-грамматические отношения двухсловных коллокаций можно представить в виде предиката  $P(x, y)$ , где  $x, y \in M$ , при этом  $x$  – главное слово словосочетания, а  $y$  – зависимое слово словосочетания. Предикат  $P(x, y) = 1$  в том случае, если сочетание семантико-грамматической информации двух рядом стоящих словоформ образует коллокацию, и  $P(x, y) = 0$ , в противном случае.

Для формального определения коллокаций выделены и описаны следующие грамматические и семантические характеристики коллокатов словосочетаний:

$$\begin{aligned}
 & a^{NNom} \vee a^{NGen} \vee a^{NAcc} \vee a^{NDat} \vee a^{Nln} \vee a^{NPr} \vee a^{VRef} \vee \\
 & \vee a^{VVonRef} \vee a^{ANom} \vee a^{AGen} \vee a^{AAcc} \vee a^{ADat} \vee a^{AIn} \vee a^{APr} = 1; \quad (1) \\
 & c^{Ag} \vee c^{Att} \vee c^{Pac} \vee c^{Adr} \vee c^{Ins} \vee c^M = 1,
 \end{aligned}$$

где использованы предметные переменные, характеризующие грамматические категории:  $a^{NNom}$  – существительное в именительном падеже,  $a^{NGen}$  – существ-

вительное в родительном падеже,  $a^{NAcc}$  – существительное в винительном падеже,  $a^{NDat}$  – существительное в дательном падеже,  $a^{Nln}$  – существительное в творительном падеже,  $a^{NPr}$  – существительное в предложном падеже,  $a^{ANom}$  – прилагательное в именительном падеже,  $a^{AGen}$  – прилагательное в родительном падеже,  $a^{AAcc}$  – прилагательное в винительном падеже,  $a^{ADat}$  – прилагательное в дательном падеже,  $a^{Aln}$  – прилагательное в творительном падеже,  $a^{APr}$  – прилагательное в предложном падеже,  $a^{VRef}$  – глагол возвратный,  $a^{VNonRef}$  – глагол невозвратный; и семантические категории:  $c^{Ag}$  – агенс,  $c^{Att}$  – атрибут,  $c^{Pac}$  – пациенс,  $c^{Adr}$  – адресат,  $c^{Ins}$  – инструмент,  $c^M$  – содержание.

Введенный на множестве словоформ  $M$  предикат  $P(x)$  обращается в 1, если главная словоформа словосочетаний обладает определенной семантико-грамматической информацией, и  $P(x) = 0$ , если главная словоформа коллокации не может обладать заданной семантико-грамматической информацией:

$$P(x) = a_x^{NNom} c_x^{Ag} \vee a_x^{NGen} c_x^{Att} \vee a_x^{NAcc} c_x^{Pac} \vee a_x^{NDat} c_x^{Adr} \vee a_x^{Nln} c_x^{Ins} \vee a_x^{NPr} c_x^M \vee a_x^{VNonRef} \quad (2)$$

Множество допустимых семантико-грамматических характеристик зависимого слова словосочетания описывается предикатом  $P(y)$ :

$$P(y) = a_y^{NNom} c_y^{Ag} \vee a_y^{NGen} c_y^{Att} \vee a_y^{NAcc} c_y^{Pac} \vee a_y^{NDat} c_y^{Adr} \vee a_y^{Nln} c_y^{Ins} \vee a_y^{NPr} c_y^M \vee a_y^{ANom} \vee a_y^{AGen} \vee a_y^{AAcc} \vee a_y^{ADat} \vee a_y^{Aln} \vee a_y^{APr} \quad (3)$$

Двухместный предикат  $P(x, y)$  описывает бинарное отношение, являющееся подмножеством конъюнкции  $P(x) \bullet P(y)$ , определяющее возможные сочетания семантико-грамматической информации словоформ двухсловных коллокаций:

$$P(x, y) = (a_y^{ANom} \vee a_y^{AGen} \vee a_y^{AAcc} \vee a_y^{ADat} \vee a_y^{Aln} \vee a_y^{APr}) (a_x^{NNom} c_x^{Ag} \vee a_x^{NGen} c_x^{Att} \vee a_x^{NAcc} c_x^{Pac} \vee a_x^{NDat} c_x^{Adr} \vee a_x^{Nln} c_x^{Ins} \vee a_x^{NPr} c_x^M) a_y^{NGen} c_y^{Att} \vee a_x^{VNonRef} a_y^{NAcc} c_y^{Pac} \quad (4)$$

Например,  $a_x^{NNom} c_x^{Ag} a_y^{NGen} c_y^{Att}$  описывает семантико-грамматически характеристики словосочетаний:

- *мова розмітки;*
- *період користування.*

Можно определить предикат семантической эквивалентности между коллокациями, определяющий семантико-грамматические характеристики коллокатов близких по смыслу словосочетаний. Отно-

шение семантической эквивалентности двух двухсловных коллокаций может быть определено как:

$$P(x_1, y_1) * P(x_2, y_2) = \gamma_1(x_1, y_1, x_2, y_2) \bullet P(x_1, y_1) \bullet P(x_2, y_2) \quad (5)$$

где знак  $*$  обозначает операцию определения смысловой близости, знак  $\bullet$  – определяет конъюнкцию, предикат  $\gamma_1(x_1, y_1, x_2, y_2)$  исключает коллокации, между которыми не может быть установлена смысловая эквивалентность.

Предикат  $\gamma_1(x_1, y_1, x_2, y_2) = a_{x_1}^{VNonRef} a_{y_1}^{NAcc} c_{y_1}^{Pac} a_{x_2}^{VNonRef} a_{y_2}^{NAcc} c_{y_2}^{Pac}$  показывает семантическую близость глагольных коллокаций. Например:

- *зберігати дані  $\approx$  тримати показники;*
- *визначати відомості  $\approx$  встановлювати дані  $\approx$  виявлення інформації.*

Предикат  $\gamma_2(x_1, y_1, x_2, y_2) = a_{x_1}^{NNom} c_{x_1}^{Ag} a_{y_1}^{NGen} c_{y_1}^{Att} a_{x_2}^{NNom} c_{x_2}^{Ag} a_{y_2}^{NGen} c_{y_2}^{Att}$  показывает семантическую близость субстантивных коллокаций, такими как:

- *набір приладдя  $\approx$  комплект устаткування;*
- *процес утворення  $\approx$  хід формування  $\approx$  процедура заснування.*

Предикат  $\gamma_3(x_1, y_1, x_2, y_2) = a_{y_1}^{ANom} a_{x_1}^{NNom} c_{x_1}^{Ag} a_{x_2}^{ANom} c_{x_2}^{Ag} a_{y_2}^{NGen} c_{y_2}^{Att}$  показывает семантическую близость между адъективной и субстантивной коллокациями, например:

- *грошовий переказ  $\approx$  відправлення коштів;*
- *інформаційний потік  $\approx$  кількість інформації.*

Таким образом, предикат семантической близости двухсловных коллокаций, у которых можно определить коллокаты как попарно семантические эквиваленты, определен как:

$$\gamma(x_1, y_1, x_2, y_2) = a_{y_1}^{ANom} a_{x_1}^{NNom} c_{x_1}^{Ag} a_{y_2}^{ANom} a_{x_2}^{NNom} c_{x_2}^{Ag} \vee (a_{x_1}^{NNom} c_{x_1}^{Ag} \vee a_{x_1}^{NGen} c_{x_1}^{Att} \vee a_{x_1}^{NAcc} c_{x_1}^{Pac} \vee a_{x_1}^{NDat} c_{x_1}^{Adr} \vee a_{x_1}^{Nln} c_{x_1}^{Ins} \vee a_{x_1}^{NPr} c_{x_1}^M) a_{y_1}^{NGen} c_{y_1}^{Att} (a_{x_2}^{NNom} c_{x_2}^{Ag} \vee a_{x_2}^{NGen} c_{x_2}^{Att} \vee a_{x_2}^{NAcc} c_{x_2}^{Pac} \vee a_{x_2}^{NDat} c_{x_2}^{Adr} \vee a_{x_2}^{Nln} c_{x_2}^{Ins} \vee a_{x_2}^{NPr} c_{x_2}^M) a_{y_2}^{NGen} c_{y_2}^{Att} \vee a_{x_1}^{VNonRef} a_{y_1}^{NAcc} c_{y_1}^{Pac} a_{x_2}^{VNonRef} a_{y_2}^{NAcc} c_{y_2}^{Pac} \quad (6)$$

В результате предикаты коллокаций, семантико-грамматические характеристики коллокатов которых обеспечат равенство единице, будут представлять близкие по смыслу или семантически эквивалентные словосочетания. А коллокации, семантико-грамматические характеристики которых обратят предикат  $\gamma_1(x_1, y_1, x_2, y_2)$  в нуль, будут обладать не близким семантическим значением.

Таким образом, логико-лингвистическая модель идентификации бинарных связей семантической эквивалентности между коллокациями позволяет формализовать отношение семантики, неявно выраженное в естественно-языковых конструкциях.

**Выводы.** В результате проведенного исследования была разработана логико-лингвистическая модель идентификации семантически близких коллокаций. Для построения модели использовался метод компонентного анализа. Для формализации семантически близких коллокаций выделены семантико-грамматические характеристики коллокатов, отношения между которыми описаны с использованием базового аппарата алгебры конечных предикатов.

Предложенная логико-лингвистическая модель призвана повысить эффективность работы существующих систем семантической обработки текстов, например, при устранении смысловой неоднозначности, извлечении фактов, построении тезауруса и др. за счет формализации как синтаксической, так и семантической информации о лексических единицах языка, и автоматизации выявления семантически эквивалентных словосочетаний.

**Список литературы:** 1. Church K. W. Word association norms, mutual information, and lexicography / K. W. Church, P. Hanks // *Computational Linguistics*, 1990. – № 16(1). – P. 22–29. 2. Evert S. Methods for the qualitative evaluation of lexical association measures / S. Evert, B. Krenn // *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, France, 2001. – P. 188–195. 3. Захаров В. П. Выделение терминологических словосочетаний из специальных текстов на основе различных мер ассоциации / В. П. Захаров, М. В. Хохлова // Интернет и современное общество «IMS-2014»: сб. научн. статей XVII всероссийской объединенной конференции, 19-20 ноября 2014 г., г. Санкт-Петербург. – СПб.: Университет ИТМО, 2014. – С. 290–293. 4. Akinina Y. S. The impact of syntactic structure on verb-noun collocation extraction / Y. S. Akinina, I. O. Kuznetsov, S. Y. Toldova //

Компьютерная лингвистика и интеллектуальные технологии : материалы международной конференции «Диалог», 29 мая - 2 июня 2013 г., г. Бекасово. – М. : Изд-во РГГУ, 2013. – № 12 (19). – Т. 1. – С. 2–17. 5. Новое в зарубежной лингвистике : Вып. XXIV. Компьютерная лингвистика / ред. Б. Ю. Городецкого. – М. : Прогресс, 1989. – 432 с. 6. Кобозева И. М. Лингвистическая семантика: Учебное пособие / И. М. Кобозева. – М. : Эдиториал УРСС, 2000. – 352 с. 7. Хайрова Н. Ф. Метод автоматической идентификации семантических корреляций терминов глоссария / Н. Ф. Хайрова, С. В. Петрасова, С. В. Ленков // Сборник научных трудов ВИКНУ имени Тараса Шевченко. – К. : ВИКНУ, 2014. – №46. – С. 128–135.

**Bibliography (transliterated):** 1. Church, K. W. "Word association norms, mutual information, and lexicography." *Computational Linguistics* 16.1 (1990): 22–29. Print. 2. Evert, S., and B. Krenn "Methods for the qualitative evaluation of lexical association measures." *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*. France, 2001. 188–195. Print. 3. Zaharov, V. P., and M. V. Hohlova "Vydelenie terminologicheskikh slovosochetaniy iz spetsial'nykh textov na osnove razlichnykh mer assotsiatsii." *Proceedings of XVII "Internet and Modern Society IMS-2014" conference*. St. Petersburg: ITMO, 2014. 290-293. Print. 4. Akinina, Y. S., I. O. Kuznetsov and S. Y. Toldova "The impact of syntactic structure on verb-noun collocation extraction." *Proceedings of the annual international Dialogue conference "Computational Linguistics and Intellectual Technologies"*. No. 12(19). Moscow: RGGU, 2013. 2–17. Print. 5. *Novoe v zarubezhnoy lingvistike: No. XXIV. Computational linguistic*. Ed. B. Yu. Gorodetskiy. Moscow: Progress, 1989. Print. 6. Kobozeva, I. M. *Lingvisticheskaya semantika*. Moscow: Editorial URSS, 2000. Print. 7. Khairova, N. F., S. V. Petrasova and S. V. Lenkov "Metod avtomaticheskoy identifikatsii semanticheskikh korrelatsiy terminov glossariya." *Collection of Scientific Papers of the Military Institute* 46 (2014): 128–135. Print.

Поступила (received) 08.07.2015

**Петрасова Світлана Валентинівна** – аспірант кафедри інтелектуальних комп'ютерних систем НТУ "ХПІ"; тел.: (093) 083-26-13; e-mail: svetapetrasova@gmail.com.

**Петрасова Светлана Валентиновна** – аспірант кафедри інтелектуальних комп'ютерних систем НТУ "ХПІ"; тел.: (093) 083-26-13; e-mail: svetapetrasova@gmail.com.

**Petrasova Svitlana Valentynivna** – Postgraduate Student at the Department of Intelligent Computer Systems, NTU "KhPI"; tel.: (093) 083-26-13; e-mail: svetapetrasova@gmail.com.

**Хайрова Ніна Феліксівна** – доктор технічних наук, професор, професор кафедри інтелектуальних комп'ютерних систем НТУ "ХПІ"; тел.: (050) 966-37-44; e-mail: nina\_khajrova@yahoo.com.

**Хайрова Нина Феликсовна** – доктор технических наук, профессор, профессор кафедры интеллектуальных компьютерных систем НТУ "ХПІ"; тел.: (050) 966-37-44; e-mail: nina\_khajrova@yahoo.com.

**Khairova Nina Feliksivna** – Doctor of Computer Linguistics, Full Professor, Professor at the Department of Intelligent Computer Systems, NTU "KhPI"; tel.: (050) 966-37-44; e-mail: nina\_khajrova@yahoo.com.