УДК 81'32

д.т.н., доц. **Хайрова Н.Ф.** (ХПИ)
**Узлов Д.Ю.** (ХПИ)

# ИДЕНТИФИКАЦИЯ КРИМИНАЛЬНО ЗНАЧИМЫХ КОЛЛОКАЦИЙ В УКРАИНОЯЗЫЧНЫХ ТЕКСТАХ

***В работе предлагается двухэтапный метод идентификации именных коллокаций в криминалистически значимых текстах украинского языка. Метод включает логико-лингвистическую модель автоматического выделения в слабоструктурированном тексте именных словосочетаний и вероятностную модель определения совместимости слов словосочетания, предназначенную для повышения точности идентификации коллокаций.***

***Ключевые слова: криминально значимые коллокации, логико-лингвистическая модель, вероятностная модель, точность идентификации.***

**Введение.** Слова, описывающие преступные деяния, имеют свою специфику и часто именно они являются индикативной признаком, по которому осуществляется отбор документов, предназначенных для последующей аналитической обработки. Понятны словосочетания *ножевое ранение*, *признаки насилия, огнестрельное ранение, взрывчатое вещество, наркотическое вещество, угон автомобиля, завладение имуществом, умышленный поджог, кража денег* и т.п. Однако, иногда интересно выявить менее привычные, но более эффективные сочетания слов для поиска криминально значимой информации, например, "*винт солянка*". Каждое из приведенных словосочетаний у профессиональных работников правоохранительной системы вызывает ассоциации с определенным видом преступления, а, следовательно, их наличие в тексте требует, по крайней мере, глубокого изучения этого текста.

В рамках семантико-синтаксического подхода коллокации (устойчивые словосочетания) рассматриваются как синтаксически связанные, лексически определённые

элементы грамматических структур, которые характеризуются семантической, синтаксической и дистрибутивной регулярностью [1] .

Как показывают проведенные исследования, в криминалистически значимых текстах особый интерес представляют именные коллокации. Поэтому на первом этапе обработки массива разнородных текстов необходимо выделить именные словосочетания, используемые в качестве объектов или характеристик данных объектов, которые определяются через взаимное информационное влияние слов в предложении. При этом анализируются только предложения, подчиняющиеся закону проективности, то есть предложения «делового стиля» [2]. Содержательный смысл условия проективности предложения состоит в том, что синтаксически связанные слова близки друг к другу и по положению в предложении. Например, именная группа может быть образована только из смежных слов. Проективность не допускает разрыва именной группы [3]. В научной и деловой украинской прозе подавляющее большинство предложений проективны.

**Постановка проблемы.** Устойчивые словосочетания (коллокации) является мощным средством в поиске криминально значимой информации. Поэтому на этапе контекстного анализа лингвистической обработки неструктурированных или слабоструктурированных массивах текстовой информации предлагается использовать двух этапный метод, базирующийся на использовании логико-лингвистической модели выделения именных грамматических словосочетаний и вероятностной модели идентификации криминалистически значимых коллокаций.

На вход контекстного анализа поступают словоформы, размеченные морфологической информацией. Поскольку на этапе морфологического анализа словоформы анализируются вне связи с контекстом, практически каждая словоформа обладает морфологической омонимией, в результате чего ей приписывается целый комплекс морфологической информации (КМИ), представляющий набор возможных морфологических альтернативных вариантов. Однако, при образовании коллокаций семантико-синтаксические связи рядом стоящих словоформ используют только определенные значения морфологических категорий [4]. Предлагается использовать математическую модель, позволяющую выделить грамматические словосочетания по типу управления и примыкания. Для повышения коэффициента точности выделяемых коллокаций используется вероятностная модель совместимости слов словосочетания в анализируемом тексте.

**Описание математической модели.** Рассмотрим множество словоформ $M = \{m_1, \dots, m_n\}$, где $n$ – количество словоформ в словаре системы. Словоформы из множества M образуют словосочетания, т.е. между словоформами устанавливаются семантико-синтаксические связи, которые можно выразить формально, используя основные средства и методы теории интеллекта [5].

Грамматическое словосочетание можно представить в виде: $m_i * m_j$ , где $m_i, m_j \in M$, а знак $*$ – обозначает, что между словоформами установлены определенные семантико-синтаксические связи.

На множестве M введем систему предикатов S так, чтобы любой предикат $P(q_m) \in S$, обращался в 1 на множестве словоформ с существующей морфологической информацией (например: *gr="S, f, inan=pl, gen"*) и был равен 0 в противном случае.

Каждому элементу *m* взаимно однозначно соответствует определенный одноместный предикат, задающий комплекс морфологической информации словоформы словосочетания. Операция соединения двух словоформ из *M*, комплексы морфологической информации которых заданы предикатами $P(q_m) \in S$ и $P(q_n) \in S$, характеризуется определенной семантико-синтаксической связью, которая определяет тип грамматического подчинения в словосочетании.

В результате семантико-синтаксического отношения двух рядом стоящих словоформ получаем множество связей между КМИ, другими словами – получаем множество возможных семантико-синтаксических связей в различных типах грамматического подчинения в словосочетаниях. В проективных предложениях украинского языка, если две

словоформы представляют именную коалицию, то связи между ними образуют два типа грамматического подчинения: согласование и управление.

Таким образом, между КМИ рядом стоящих словоформ предложения существует бинарное отношение, являющееся подмножеством декартового произведения этих комплексов, которое можно представить с помощью некоторого двухместного предиката:

$$P(q_m, q_n) \rightarrow P(q_m) \bullet P(q_n), \quad (1)$$

где $\bullet$ – операция конъюнкции предикатов.

Возможность согласования морфологических информаций не зависит от того, к каким словоформам они относятся. На декартовом произведении множества $S * S$ зададим предикат $\gamma_\iota(q_m, q_n)$, принимающий значение 1, если морфологические информации словоформ $q_m$ и $q_n$ связаны грамматической связью согласования или управления, и значение 0 в противном случае. Практически никогда подмножество согласуемой морфологической информации не совпадает с декартовым произведением всех возможных связей, поэтому морфологическая информация тех рядом стоящих словоформ, которые не согласуются в данном типе грамматического подчинения, исключаются из формулы (1*)* множителем $\gamma_\iota(q_m, q_n)$, $\iota = 1,2$ (согласование, управление). Таким образом, бинарное отношение на множестве рядом стоящих словоформ предложения для возможных словосочетаний украинского языка задается формулой:

$$P(q_m) * P(q_n) = \gamma_\iota(q_m, q_n) \bullet P(q_m) \bullet P(q_n), \quad (2)$$

где знак $*$ обозначает операцию соединения комплексов морфологических информации словоформ, образующих именное словосочетание (операцию связи МИ двух рядом стоящих словоформ, т.е. знак $*$ указывает на то, что две рядом стоящие словоформы связаны между собой семантико-синтаксической связью). Действительно, логическое произведение предикатов $P(q_m)$ и $P(q_n)$ описывает всевозможные связи КМИ между двумя рядом стоящими словоформами в предложении, а предикат $\gamma_\iota(q_m, q_n)$ исключает часть связей, которые не реализуются в данном типе грамматического подчинения украинского языка.

**Описание используемого метода.** Рассмотрим работу данной модели для извлечения именных словосочетаний из текстов украинского языка. Выберем простейшую систему морфологических категорий и их значений, состоящую из части речи и наиболее существенных морфологических признаков. В случае использования более сложного морфо анализатора, систему грамматических категорий можно расширить без изменения алгоритма.

Морфологическая информация ключевой (свободной) словоформы словосочетания [6] может быть представлена предметной переменной $x$, область изменения которой:

$$x^S \vee x^{So} \vee x^V \vee x^{Vp} \vee x^{PR} \vee x^A \vee x^{ADV} \vee x^{AN} \vee x^N \vee x^{Ly} = 1,$$

где $x^S$ – существительное именительного падежа, $x^{So}$ – существительное косвенного падежа, $x^V$ – глагол не прошедшего времени, $x^{Vp}$ – глагол прошедшего времени, $x^{PR}$ – предлог, $x^A$ – прилагательное, $x^{ADV}$ – причастие, $x^{AN}$ – порядковое числительное, $x^N$ – количественное числительное, $x^{Ly}$ – служебная часть речи.

Морфологическая информация несвободной компоненты двухсловного словосочетания украинского языка может быть представлена предметной переменной $y$, область изменения которой аналогична

$$y^S \vee y^{So} \vee y^V \vee y^{Vp} \vee y^{PR} \vee y^A \vee y^{ADV} \vee y^{AN} \vee y^N \vee y^{Ly} = 1,$$

где $y^S$– существительное именительного падежа, $y^{So}$ – существительное косвенного падежа, $y^V$ – глагол не прошедшего времени, $y^{Vp}$ – глагол прошедшего времени, $y^{PR}$ – предлог, $y^A$ – прилагательное, $y^{ADV}$ – причастие, $y^{AN}$ – порядковое числительное, $y^N$ – количественное числительное, $y^{Ly}$ – служебная часть речи.

При грамматическом подчинении по типу согласования (*умышленный поджог*, *наркотическим веществом*, *украденный пистолет*, *произведенной взрывчаткой*, *второго срока* и др.) предикат $\gamma_1(q_n, q_m)$ формулы (2) может быть представлен следующим образом:

$$\gamma_1(q_n, q_m) = x^S y^A \vee x^{So} y^A \vee x^S y^{ADV} \vee x^{So} y^{ADV} \vee x^{So}\, y^{AN} \vee x^S y^{AN}.$$

При грамматическом подчинении по типу управления (*удар топором*, *ударом ножа*, *на рукоятке*, *угнать автомобиль*, *украл автомобиль* и др.) предикат $\gamma_2(q_n, q_m)$ формулы (2) будет представлен формулой:

$$\gamma_2(q_n, q_m) = x^S y^{So} \vee x^{So} y^{So} \vee x^{PR} y^{So} \vee x^V y^{So} \vee x^{Vp} y^{So}.$$

Тогда в соответствии с формулой (2) множество возможных связей комплексов морфологической информации в словосочетаниях по типу согласования описывается предикатом $P_1(q_m, q_n)$:

$$P_1(q_m, q_n) = (x^S y^A \vee x^{So} y^A \vee x^S y^{ADV} \vee x^{So} y^{ADV} \vee x^{So}\, y^{AN} \vee x^S y^{AN})\,(x^S \vee x^{So} \vee x^V \vee x^{Vp} \vee x^{PR} \vee x^A \vee x^{ADV} \vee x^{AN} \vee x^N \vee x^{Ly})\,(y^S \vee y^{So} \vee y^V \vee y^{Vp} \vee y^{PR} \vee y^A \vee y^{ADV} \vee y^{AN} \vee y^N \vee y^{Ly}). \quad (3)$$

Множество возможных связей комплексов морфологической информации в словосочетаниях по типу управления, задаваемое с помощью предиката $P_2(q_m, q_n)$ представлено формулой:

$$P_2(q_m, q_n) = (x^S y^{So} \vee x^{So} y^{So} \vee x^{PR} y^{So} \vee x^V y^{So} \vee x^{Vp} y^{So})\,(x^S \vee x^{So} \vee x^V \vee x^{Vp} \vee x^{PR} \vee x^A \vee x^{ADV} \vee x^{AN} \vee x^N \vee x^{Ly})\,(y^S \vee y^{So} \vee y^V \vee y^{Vp} \vee y^{PR} \vee y^A \vee y^{ADV} \vee y^{AN} \vee y^N \vee y^{Ly}). \quad (4)$$

При подстановке КМИ словоформ словосочетания, полученных на этапе морфологического анализа, в формулы (3 – 4), предикаты, которые описывали тип словосочетания, не присущий данным словоформам, обращаются в нуль.

Предикаты, которые принимают значение 1, описывают словосочетания согласования и управления. Данные бинарные предикаты позволяют получить ограниченное множество грамматических словосочетаний, являющихся потенциальными коллокациями в слабоструктурированных криминально значимых текстах украинского языка.

### 1. Использование вероятностного похода для повышения эффективности

Для повышения точности определения устойчивых криминальных словосочетаний текста на втором этапе контекстного анализа определяется вероятностная совместимость слов словосочетания в анализируемом тексте:

$$P(w_1, w_2) = \log_2\left(\frac{P(w_1 \wedge w_2)}{P(w_1) * P(w_2)}\right), \quad 5)$$

где $P(w_1,w_2)$ – вероятность данной двухсловной коллокации в тексте, $P(w_1 \wedge w_1)$ – вероятность совместного появления слов, $P(w_1)$, $P(w_2)$ – вероятность независимого появления слов в рассматриваемом тексте или массиве текстов. Если $P(w_1,w_2) > 0$, то считаем, что данный набор нормализованных слов представляет коллокацию; в противном случае, если $P(w_1,w_2) < 0$, то данный набор представляет изолированные слова, не являющиеся коллокациями криминально значимой текстовой информации.

**Выводы.** Таким образом, предложенный в работе метод выделения именных коллокаций из текстов украинского языка позволяет на этапе контекстного анализа выделить индикативные признаки криминалистически значимых текстов, представляющих последующий интерес для аналитической обработки. Используемая логико-лингвистическая

модель позволяет выделить грамматические словосочетания по типу управления и согласования, с существительным в качестве ключевого слова. Для повышения точности определения криминально значимых коллокаций используется вероятностная совместимость слов словосочетаний в анализируемом тексте.

ЛІТЕРАТУРА:

1. Carnie A. Syntax: A Generative Introduction (Introducing Linguistics). 2nd Edition Oxford, П.И.Браславский. Тезаурус для расширения запросов к машинам поиска Интернета: структура и функции. Режим доступу: http://www.dialog-21.ru/Archive/2003/Braslavskij.htm .
2. Гладкий А. В. Грамматики деревьев: опыт формализации преобразований синтаксических структур естественного языка / А.В. Гладкий И. А. Мельчук // Информ. вопросы семиотики, лингвистики и автоматического перевода. – 1971. – Вып. 1. – С. 16-41.
3. Попов Э.В. Общение с ЭВМ на естественном языке. – М.: Наука, 1982.—360с.
4. Ситников Д.Э., Шаронова Н.В., Хайрова Н.Ф. Моделирование семантико-синтаксических отношений грамматических словосочетаний // Пробл. бионики. – 1999. – Вып.50. – С. 179-184.
5. Бондаренко М. Ф. Теория интеллекта: учебник/ Бондаренко М. Ф., Шабанов-Кушнаренко Ю. П. Харьков: Комп. СМИТ, 2007. – 576 с.
6. Мельчук И.А. Русский язык в модели "Смысл-текст": Вен. слав. Альманах. – М., Вена: Шк. "Языки русской культуры". – 1995. – Вып. XXVIII. – 682с.

**Рецензент:** **д.т.н., проф. Шворов С.А.,** Національний університет біоресурсів і природокористування України

**д.т.н., доц. Хайрова Н.Ф., Узлов Д.Ю.**

## ІДЕНТИФІКАЦІЯ КРИМІНАЛЬНО ЗНАЧИМИХ КОЛЛОКАЦІЙ В УКРАЇНСЬКОМОВНИХ ТЕКСТАХ

***У роботі пропонується двоетапний метод ідентифікації іменних колокацій в криміналістично значущих текстах української мови. Метод містить логіко-лінгвістичну модель автоматичного виділення в слабко структурованому тексті іменних словосполучень і вірогідну модель визначення сумісності слів словосполучення, що призначена для підвищення точності ідентифікації колокацій.***

***Ключові слова: кримінально значущі колокації, логіко-лінгвістична модель, імовірнісна модель, точність ідентифікації.***

**N. Hayrova, D. Uzlov.**

## IDENTIFICATION OF CRIMINAL SIGNIFICANT COLLOCATION IN THE UKRAINIAN-TEXTS

***This paper proposes a two-step method for identifying noun collocations in criminal meaningful texts in English. The method involves the logical-linguistic model for automatic selection noun phrases in semistructured text and probabilistic model for determining the compatibility of words phrases. The last model is intended to improve the accuracy of collocations identification.***

***Keywords: criminal meaningful collocations, logical-linguistic model, probabilistic model, identification accuracy***