Извлечение коллокаций из корпуса украинских текстов

Kolokacijų nustatymas ukrainiečių kalbos tekstuose

SAL 27/2015

Извлечение коллокаций из корпуса украинских текстов

Received 04/2015

Accepted 08/2015

COMPUTATIONAL LINGUISTICS / KOMPIUTERINĖ LINGVISTIKA

Татьяна Бобкова

Кандидат филологических наук, доцент Киевского национального лингвистического университета, Украина.

http://dx.doi.org/10.5755/j01.sal.0.27.13747

Аннотация

В статье описывается методика извлечения двусловных коллокаций из корпуса украинских законодательных текстов. Существующие методики выделения коллокаций основываются на подходах, отличающихся критериями идентификации и последовательностью применяемых процедур. В работе обосновывается необходимость использования корпусно-ориентированного подхода, основанного на идентификации коллокации как статистически значимой единицы и применении корпусных методов обработки текстов. Коллокация определяется как неслучайное сочетание двух слов, регулярно встречающихся вместе, и характерное как для текстов определенного функционального стиля, так и для языка в целом. Разработанная методика идентификации двусловных коллокаций, позволяет на основе статистической обработки и использования программ лемматизации автоматически извлекать устойчивые двухсловные сочетания из подкорпуса украинских текстов. Результаты извлечения нуждаются в последующем редактировании с целью снятия омонимии и определения грамматически правильных коллокаций. Повышение эффективности результатов автоматического формирования списка обеспечит применение большего по объему корпуса текстов и лингвистических фильтров идентификации коллокаций.

КЛЮЧЕВЫЕ СЛОВА: коллокация, текст, корпус, извлечение, идентификация, лингвистический подход, корпусно-ориентированный подход.

Введение

Идентификация и извлечение коллокаций является частью общей проблемы распознавания естественно-языкового текста. Актуальность подобных исследований объясняется, прежде всего, высокой частотностью коллокаций в текстах разных функциональных стилей (Fontenelle, 1994; Marcinkevičienė, 2010; Seretan, 2011; Sinclair, 1991; Smadja & McKeown, 1990; Tognini-Bonelli, 2000; Большакова, 2011; Зацеркляний, 2013; Хохлова, 2010; Шкурко, 2012; Ягунова & Пивоварова, 2011). Результаты исследования коллокаций находят широкое применение при разработке систем информационного поиска, автоматического анализа текста, машинного перевода, составления словарей, тезаурусов, баз данных и т.д. В частности, необходимость решения прикладных задач по автоматическому анализу украинского текста привела к появлению ряда исследований, посвященных проблемам распознавания и извлечения коллокаций (Дарчук, 2013; Шкурко, 2012), в том числе и в области информационного поиска (Зацеркляний, 2013; Хайрова & Узлов, 2013).

Research Journal
Studies about Languages
No. 27/2015
ISSN 1648-2824 (print)
ISSN 2029-7203 (online)
pp. 93-105
DOI 10.5755/j01.sal.0.27.13747

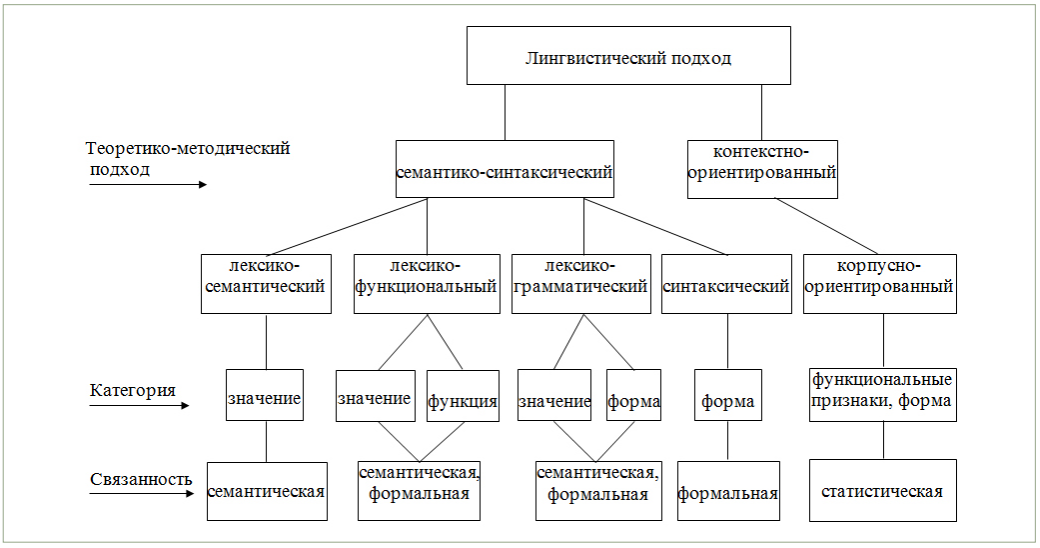
Несмотря на применение различных подходов к выделению коллокаций, основные проблемы отбора составляют: установление критериев идентификации (McKeown & Radev, 2000, p. 509; Seljan & Gašpar, 2012, p. 152; Smadja & McKeown, 1990, p. 258; Лендау, 2012, с. 304; Романюк, 2011, с. 158), классификация коллокаций (McKeown & Radev, 2000, p. 511; Smadja & McKeown, 1990, p. 258) и оценивание эффективности используемых приемов и процедур (Seljan & Gašpar, 2012, p. 149; Лендау, 2012, с. 303; Червяк, 2011, с. 46). Проблема состоит в том, что ни одна система не извлекает весь диапазон коллокаций из анализируемого текста (Smadja & McKeown, 1990, p. 258). Принципы выделения коллокаций зачастую ограничены традициями определенной школы, интуицией исследователя или узко заданной темой (Ягунова & Пивоварова, 2011). В связи с этим возникает необходимость усовершенствования методики распознавания коллокаций в естественно-языковом тексте на основе объективных критериев. Цель данной статьи – описать методические этапы извлечения коллокаций из корпуса украинских текстов.

Основные подходы к идентификации и извлечению коллокаций

Под извлечением коллокаций понимается операция, на вход которой подаются анализируемые тексты, а на выходе – результат их обработки – составленный список потенциальных коллокаций (Seljan & Gašpar, 2012, p. 150). Соответственно, автоматическое извлечение коллокаций предполагает использование коллекции текстов со специальным лингвистическим инструментарием (Ananiadou, 1994, p. 1034; Seljan & Gašpar, 2012, p. 156). Интерес со стороны специалистов разных областей к исследованию троякой природы феномена способствовал формированию различных подходов к выделению коллокаций, среди которых выделяют лингвистический и нелингвистический – статистический, математический (Ananiadou, 1994; McKeown & Radev, 2000; Seljan & Gašpar, 2012; Smadja & McKeown, 1990; Wermter, 2008; Романюк, 2011; Хохлова, 2010; Ягунова & Пивоварова, 2011). Указанные подходы дифференцируются по критериям идентификации коллокаций и последовательности процедур, применяемых для формирования списка (Ananiadou, 1994, p. 1034; Wermter, 2008, pp. 53–54). В теоретико-методическом аспекте лингвистический подход к выделению коллокаций опирается на принципы семантико-синтаксического и контекстно-ориентированного (Рис. 1).

В зависимости от концепции и применяемой методики выделения коллокаций в рамках семантико-синтаксического подхода следует выделить: лексико-семантический, лексико-функциональный, лексико-грамматический и синтаксический (Бобкова, 2014, с. 16–18).

Рис 1
Типология подходов к выделению коллокаций



Семантико-синтаксический подход: методики выделения коллокаций

Намеченным выше подходам идентификации и экстракции коллокаций из текста соответствуют методики (полу)закрытого и открытого списков (Ягунова & Пивоварова, 2011). Методика закрытого списка предполагает наличие заранее заданных потенциальных ответов системы, или реестра искомых единиц, в данном случае – устойчивых сочетаний. Использование более сложного в методическом отношении полузакрытого списка основывается на закрытом списке значений признака или набора признаков искомых единиц. Методика открытого списка теоретически не предполагает никаких, заранее заданных ограничений в интерпретации статистическоих результатов обработки текста (Ягунова & Пивоварова, 2011). Соответственно, в традиционных лингвистических исследованиях для выделения коллокаций используется в основном методика (полу)закрытого списка, а в корпусно-ориентированных – методика открытого списка.

Семантико-синтаксический подход предусматривает анализ коллокаций в аспекте теории устойчивых сочетаний (Ш. Балли, В.В. Виноградов, Н.П. Дарчук, В.Н. Телия, Н.М. Шанский, Д.Н. Шмелев) и грамматики конструкций (Ч.Дж. Филмор). Коллокации рассматриваются как комплексные семантико-синтаксические единицы, характеризующиеся семантической, синтаксической и дистрибутивной регулярностью. Связанность чаще всего определяется семантической совместимостью элементов устойчивого сочетания (Kapstad, 2006, с. 19–20; Борисова, 1995, с. 77; Киверник, 2012) и/или определенной синтаксической моделью (Seretan, 2011, p. 27; Хохлова, 2010, с. 3). Наиболее распространенным является лексико-семантический подход, поскольку в лингвистических исследованиях в определении коллокации чаще всего постулируется "лексический приоритет" (Marcinkevičienė, 2010, s. 88). В аспекте лексической сочетаемости коллокации понимаются как определенный тип устойчивого сочетания, связанность которого обусловлена семантическими отношениями. Базовый (связанный) компонент является аргументом, а свободный – коллокатом, функцией от данного аргумента (Левицкий, 2006, с. 192; Хохлова, 2010, с. 4–5). В качестве основных признаков коллокации выделяют идиоматичность, устойчивость употребления, воспроизводимость в речи, невозможность выведения значения сочетания из значений компонентов и наличие самостоятельного значения коллоката (Kapstad, 2006, с. 19–20; Большакова, 2011, с. 78).

К описанному выше подходу примыкает лексико-функциональный подход теории "Смысл↔Текст" (Ю.Д. Апресян, А.К. Жолковский, И.А. Мельчук, Л.Н. Иорданская), также предусматривающий анализ коллокаций как подкласса несвободных сочетаний, или фразем. Коллокация определяется как сочетание семантической доминанты, определяющей выбор другого компонента для передачи значения всего словосочетания. Лексическая функция ставит в соответствие определенному слову или словосочетанию X другое слово, или словосочетание Y, связанное с X по смыслу (Апресян, 1995, с. 43). Собственно, это функция в математическом смысле: главная лексема X является ее аргументом, а множество возможных (для X) выражений смысла f, то есть Y – ее значением: если X = *преступление*, смысл f = '*делать*' выражается через Y = *совершать*, а если X = *лекция*, то идентичный смысл выражается через Y = *читать* (Kapstad, 2006, с. 28). Теория лексической функции широко используется в современных прикладных исследованиях для описания значения, синтаксической, актантной структуры коллокаций и определения функциональной зависимости коллоката (Fontenelle, 1994, p. 46; McKeown & Radev, 2000, p. 511; Wermter, 2008, pp. 12–15, 57–58). Но, как показывает анализ исследований (Kapstad, 2006, с. 28–31; Киверник, 2012), трудно поддающиеся формализации семантические характеристики вряд ли могут служить дифференциальными признаками коллокаций в естественно-языковом тексте.

В соответствии с лексико-грамматическим подходом (Е.И. Большакова, Е.Г. Борисова, О.А. Митрофанова) коллокации рассматриваются как семантико-синтаксические единицы,

или лексически определенные элементы грамматических структур. При этом коллокации часто отождествляются с синтаксическими конструкциями и определяются как "последовательности слов, совместное употребление которых значительно превышает случайный уровень" (Большакова, 2011, с. 22). Данный подход предполагает извлечение коллокаций с помощью установленных лексико-грамматических шаблонов – структурных моделей языковых конструкций с указанием существенных грамматических признаков лексем, входящих в состав языковых выражений определенного класса, и синтаксических условий их употребления: *атрибутивные* – N+N2, *объектные* – V+N2, *компаративные отношения* – N+Adjcomp+N2 (Хохлова, 2010, с. 10–11, с. 19–20). Подобно этому синтаксический подход (V. Seretan, Е.И. Большакова, В.А. Гладка; В.В. Шкурко) основывается на интерпретации коллокации как синтагматически ограниченной лексической сочетаемости синтаксически связанных элементов и опирается в основном на грамматику конструкций Ч.Дж. Филмора. При этом во внимание принимаются два основных признака коллокаций: соблюдение определенного порядка следования компонентов и отсутствие синтаксических трансформаций (Гладка, 2013, с. 17). Коллокация определяется как последовательность слов, образующая синтаксическую конструкцию – *существительное + прилагательное, глагол + существительное (объект), существительное + предлог + существительное* (Seretan, 2011, p. 13; Гладка, 2013, с. 17) или идиоматическое сочетание, семантика которого шире семантики составляющих (Ягунова & Пивоварова, 2011). В отличие от лексико-семантического и лексико-функционального в основу синтаксического подхода к выделению коллокаций положены критерии синтаксической связанности, контактного расположения в тексте и наличия подчинительной связи между компонентами (Шкурко, 2012, с. 31). Полученные таким образом результаты экстракции коллокаций широко используются в разработке современных лексикографических систем (Seretan, 2011, p. 28; Шкурко, 2012, с. 32) и автоматического анализа текста (Seretan, 2011, p. 28; Дарчук, 2013, с. 119; Хайрова & Узлов, 2013, с. 147; Шкурко, 2012, с. 32).

В методическом отношении семантико-синтаксический подход к выделению коллокаций основываются на использовании методик закрытого и полузакрытого списков. Процедура идентификации и извлечения коллокаций ограничивается заданным списком устойчивых сочетаний (или их признаков), составленным на основе имеющихся словарей, реестров лингвистов-экспертов (Seljan & Gašpar, 2012, p. 153; Большакова, 2011, с. 120; Ягунова & Пивоварова, 2011), баз терминов и корпусов. Так, в системе автоматического грамматического анализа украинского текста (АГАТ) коллокации определяются как эквивалентные слову несвободные сочетания и включаются в словник электронного словаря фразеологизмов наряду с бесспорными идиомами: *приймати участь, діаметрально протилежний* (Дарчук, 2013, с. 120). Учет принадлежности коллокации определенному домену способствует широкому применению данной методики в изучении предметных областей (McKeown & Radev, 2000, p. 510; Seljan & Gašpar, 2012, pp. 152, 156). В частности, на основе закрытого списка реализовано поиск контекстов употребления терминологических сочетаний в корпусе текстов по компьютерной лингвистике (Bobkova, 2009, p. 39). Более сложным в методическом отношении представляется извлечение коллокаций, не включенных в закрытый список терминов – *approve draft terms, enter into force* и др. (Seljan & Gašpar, 2012, p. 152), или на основе полузакрытого списка.

Методика полузакрытого списка предполагает установление ограничений на извлечение коллокаций по формальным и качественным признакам. В качестве таких признаков используются лексические варианты (Ягунова & Пивоварова, 2011), функциональные модели (McKeown & Radev, 2000, p. 511; Wermter, 2008, pp. 57–58), синтаксические конструкции (Большакова, 2011, с. 22; Шкурко, 2012, с. 33; Ягунова & Пивоварова, 2011)

и морфолого-синтаксические модели (Хайрова & Узлов, 2013, с. 147; Хохлова, 2010, с. 3; Шкурко, 2012, с. 33). В частности, на принципах лексико-грамматического и синтаксического подходов построена идентификация субстантивно-адъективных и субстантивно-субстантивных коллокаций в украинских криминально значимых текстах (Хайрова & Узлов, 2013, с. 147), а также лексикографическая система для экстракции субстантивных, адъективных и глагольных коллокаций из законодательных текстов (Шкурко, 2012, с. 32–33). К основным недостаткам семантико-синтаксического подхода к выделению коллокаций следует отнести ресурсоемкость (Seljan & Gašpar, 2012, p. 156; Smadja & McKeown, 1990, p. 252; Червяк, 2011, с. 41), субъективность критериев идентификации, ограниченность диапазона извлекаемых коллокаций (Seljan & Gašpar, 2012, p. 152; Smadja & McKeown, 1990, p. 258), в том числе и в аспекте анализа морфологических и функциональных характеристик (Seljan & Gašpar, 2012, p. 153). Перспективы установления объективных признаков и снижения ресурсоемкости процесса за счет автоматизации экстракции коллокаций появляются с внедрением корпусно-ориентированного подхода.

Корпусно-ориентированный подход: концепция и методика выделения коллокаций

В основу формирования концепции корпусного подхода к исследованию коллокаций (R. Marcinkevičienė, J. Sinclair, E. Tognini-Bonelli, М.В. Хохлова, Е.В. Ягунова, Л.М. Пивоварова) положены идеи и эмпирические традиции британского контекстуализма Дж. Фьорза. Основным достижением фьорзианства считается (Tognini-Bonelli, 2000, p. 209) определение двух взаимосвязанных формальных признаков контекста: 1) коллокации – сопровождения слова, или словесного материала, в котором оно чаще всего встречается, и 2) коллигации – взаимосвязи грамматических категорий в синтаксической конструкции. К фьорзианскому подходу концептуально близка теория элементарных полей В. Порцига (Marcinkevičienė, 2010, s. 71; Левицкий, 2006, с. 202–206), развиваемая в украинской семасиологии В.В. Левицким. В частности, применение квантитативных методов для изучения элеменарных полей позволило установить, что переход от свободных словосочетаний к устойчивым имеет "градуированный характер", а высокая степень связи между компонентами является неслучайной (Левицкий, 2006, с. 205–206). Однако, несмотря на осознание перспектив изучения коллокаций, реальные возможности получения и верификации данных появились благодаря использованию электронных корпусов текстов.

Подобно контекстно-ориентированному корпусный подход изначально определяется опорой на контекст, понимаемый как коллекция текстов или единичные тексты коллекции (Большакова, 2011, с. 23; Ягунова & Пивоварова, 2011). Соответственно, коллокация определяется как сочетание двух и более слов, встречающееся вместе чаще, чем можно было бы ожидать случайно (Sinclair, 1991; Большакова, 2011; Хохлова, 2010; Ягунова & Пивоварова, 2011) и характерное как для языка в целом, так и для текстов определенного типа или даже для (под)выборки текстов (Большакова, 2011, с. 23; Ягунова & Пивоварова, 2011). Основы теоретических инноваций в концепции коллокации Дж. Фьорза были заложены роботами Дж. Синклера по переопределению единицы значения в аспекте корпусных данных (Tognini-Bonelli, 2000, p. 214). Феномен коллокации рассматривается как проявление принципа идиоматичности – наличия в языке большого количества готовых фраз или их полуфабрикатов, используемых носителем языка по выбору (Marcinkevičienė, 2010, p. 60; Sinclair, 1991, p. 110). На основе обобщения корпусных данных Дж. Синклером (Sinclair, 1991, pp. 112–113) установлены характерные признаки коллокаций: 1) неопределенность объема значения лексических единиц (*set eyes on*); 2) возможность лексической (*in some cases/in some instances*) и синтаксической вариативности (*not/hardly/scarcely, recriminate isn't in his nature/isn't in his nature of an academic*); 3) закономерность появления высокочастотных слов (*hard work, hard evidence*) и 4) тенденция к

совместной встречаемости слов в определенных грамматических конструкциях (*set about testing*) и в определенном семантическом окружении (*happen – accident*).

Вариативность лингвистических признаков вызывает необходимость в описании коллокаций апеллировать нетрадиционными категориями (Tognini-Bonelli, 2000, p. 206), т. е. функциональными и статистическими характеристиками (см. Рис.1). В отличие от функции как системного признака единицы языка, функциональные признаки определяются как совокупность характеристик, приобретенных единицей в речи – устном или письменном тексте (Перебийніс & Бобкова, 2008 с. 446). К функциональным признакам относят степень употребительности (частоту), сочетаемость с другими единицами, позицию в речевой последовательности, степень реализации системных признаков (в частности, словоизменительных форм), коммуникативное назначение, прагматическую или эмотивную нагрузка, стилистическую окраску (Перебийніс & Бобкова, 2008 с. 446). В узком смысле под функциональными признаками коллокации понимаются статистические характеристики, что является основанием для определения коллокации как "сугубо статистического феномена" (Fontenelle, 1994, p. 47) и классификации корпусно-ориентированного подхода как статистического (Гладка, 2013, с. 15; Хохлова, 2010, с. 4–5; Ягунова & Пивоварова, 2011). Действительно, в аспекте используемых методов анализа корпусно-ориентированый подход является нелингвистическим и характеризуется перемещением внимания с лингвистических проблем фразеологизации на повышение эффективности автоматического выявления устойчивых сочетаний в текстах (Зацеркляний, 2013, с. 184; Романюк, 2011, с. 158–159). Тем не менее, статистический подход к извлечению коллокаций существенно отличается от корпусно-ориентированного.

Статистический подход к выделению коллокаций основывается на замене лингвистических критериев идентификации статистическими (Зацеркляний 2013, с. 184–186; Романюк, 2011, с. 163; Червяк, 2010, с. 42) или математическими (Хайрова & Узлов, 2013, с. 148) и в соответствие с методикой открытого списка предполагает обработку всего текстового материала без заранее заданных ограничений (Ягунова & Пивоварова, 2011). Указанный подход обеспечивает снижение ресурсоемкости процесса экстракции, большее покрытие текста (Ananiadou, 1994, p. 2035), преимущества в изучении вариативности словоизменительных форм и синтаксических конструкций (Fontenelle, 1994, p. 47; Seljan & Gašpar, 2012, pp. 151–152) и выявление отсутствующих в закрытых списках и словарях коллокаций (Лендау, 2012, с. 303; Червяк, 2010, с. 46). Однако проблема состоит в том, что в автоматически составленных реестрах коллокаций возрастает процент ошибок идентификации. В частности, по результатам обработки параллельного англо-хорватского корпуса больше половины выделенных устойчивых сочетаний составили "фальшивые примитивы" й повторы, требующие обязательного постредактирования (Seljan & Gašpar, 2012, p. 152). Кроме того, используемые статистические меры разработаны только для двусловных сочетаний и не способны дифференцировать ядро и коллокат (Durrant, 2010, pp. 131–132). Таким образом, троякая природа коллокаций не позволяет полностью отказаться от "символических методов" (Романюк, 2011, с. 158) обработки естественно-языкового текста.

Возникает необходимость внедрения подхода, совмещающего объективное описание статистических и лингвистических признаков коллокаций (Seljan & Gašpar, 2012; Романюк, 2011). Прототипом корпусно-ориентированного подхода послужила методика Ф. Смаджа (Smadja & McKeown, 1990), разработанная для системы Xtract и нарушившая "канонический порядок анализа" (Wermter, 2008, p. 54), используемый в лингвистических подходах к выделению коллокаций. В соответствие с данным подходом в качестве исходных данных для идентификации коллокаций в тексте рассматриваются не лингвистические, а статистические характеристики (Wermter, 2008, p. 54). Методика Ф. Смаджа существенно отличается от лингвистических и статистического подходов поэтапным

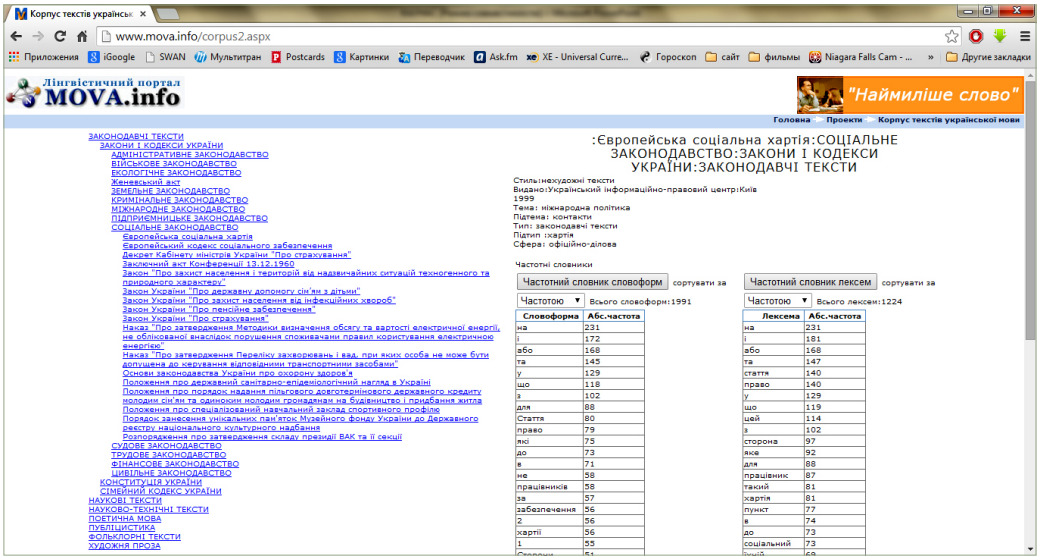
применением статистического анализа, трех лингвистических фильтров (позиционного, синтаксического, морфологического) и постредактированием результатов экспертом-лексикографом (Smadja & McKeown, 1990, p. 253). Таким образом, существенным преимуществом корпусно-ориентированного подхода является использование статистико-лингвистического аппарата корпуса для выявления релевантных грамматически правильных и семантически значимых коллокаций. Развитие описанного корпусно-ориентированного подхода к выделению коллокаций и повышение эффективности результатов обеспечивается использованием больших объемов корпусов текстов, встроенных программ лемматизации, морфолого-синтаксических фильтров (Seljan & Gašpar, 2012, p. 156) и латентного семантического индексирования (Романюк, 2011, с. 163). Кроме того, разработанные для анализа определенного подъязыка программы и алгоритмы применимы для описания разных предметных областей (McKeown & Radev, 2000, p. 507) и функциональных стилей. На современном этапе внедрение корпусно-ориентированного подхода к выделению коллокаций ограничивается в основном из-за несовершенства лингвистических процессоров, поэтому результаты автоматической обработки текстов нуждаются в обязательном постредактировании (Seljan & Gašpar, 2012, p. 156; Smadja & McKeown, 1990, p. 253). Описанные выше принципы корпусно-ориентированного подхода были положены в основу составления Словаря коллокаций украинского юридического дискурса (Бобкова, 2015, s. 44). С целью автоматического составления списка коллокаций в рамках Корпуса украинского языка был составлен подкорпус законодательных текстов (Законодавчі тексти, 2014).

Составление подкорпуса законодательных текстов

На данном этапе программные средства Корпуса украинского языка не предоставляют возможности автоматического формирования списков коллокаций: лингвистически-поисковый аппарат корпуса позволяет осуществлять только поиск отдельных словосочетаний с помощью заданных лемм и определенных грамматических ограничений (Kotsyba, 2013). Стандартный набор корпус-менеджера, включающий программы морфологического кодирования, лемматизации, поиска по словоформам, леммам и грамматическим кодам, требует дополнения соответствующим программным обеспечением, позволяющим автоматически извлекать из корпуса коллокации и получать информацию относительно их статистических и лингвистических характеристик. В соответствии с предлагаемой в данном исследовании методикой автоматическое составление списка коллокаций основывается на лексикографической базе подкорпуса законодательных текстов (Рис. 2), использовании поискового аппарата Корпуса украинского языка и специально разработанного программного обеспечения. В основу составления подкорпуса законодательных текстов, изначально планировавшегося для составления Словаря коллокаций украинского юридического дискурса, положены требования, предъявляемые к корпусным объектам, а именно: репрезентативность, сбалансированность лингвистического материала, ограниченность объема, стандартность и исследовательское предназначение (Бобкова, 2015, s. 37–39).

Репрезентативность (Лендау, 2012, с. 325) подкорпуса законодательных текстов обеспечивается: документальной емпирической базою, диапазоном текстовых типов, общим объемом подкорпуса (свыше 1 млн. слов) и временным интервалом создания включенных текстов. Прежде всего, репрезентативность составленного для словаря коллокаций подкорпуса обеспечивается аутентичностью текстов документального источника – "Собрания законодательства Украины" (Омега, 2009). Выбор в качестве лексикографической базы официальных документов определяется максимальной насыщенностью в данных текстах различных устойчивых сочетаний (Marcinkevičienė, 2010, pp. 61–62; Борисова, 1995, с. 17): неоднословных наименований (*Паризька конвенція, асамблея союзу*), терминологических сочетаний (*міжнародна реєстрація, юридична особа*), речевых формул

Рис. 2
Подкорпус законодательных текстов



и клише (*в установленому порядку, додержуватись положень*), составных предлогов и союзов (*у відповідності з, у силу того, що*).

На начальном этапе планирования подкорпуса для соблюдения принципа сбалансированности было принято решение о максимальном представлении всех текстовых типов законодательных документов. Однако при отборе было выявлено, что тексты документов характеризуются различной длиной, колеблющейся в пределах от 2 слов (*Цілком таємно*) до 44,020 тыс. слов, поэтому для достижения сбалансированности подкорпуса было решено использовать диапазон текстовых типов, или жанров, представленных наибольшими по объему текстами. В настоящее время общий объем подкорпуса законодательных текстов (1.157 млн. слов) представлен 43 текстовыми типами 199 образцами текстов. В соответствии с конечной целью исследования – составлением словаря коллокаций при отборе в подкорпус предпочтение отдавалось исключительно полнотекстовым документам, что обеспечило их структурную и лексическую завершенность. Подкорпус законодательных текстов является однородным с точки зрения языка, функционального стиля и хронологии отобранных текстов (с 1992 г. – по настоящее время). Общий объем собранных текстов обеспечивает достоверность результатов анализа грамматических конструкций (Renouf, 2007, p. 29) и частотных лексических единиц (Marcinkevičienė, 2010, s. 14), что соответствует исследованию коллокаций. Таким образом, отобранные в соответствие с эксплицитными и имплицитными критериями тексты законодательных документов обеспечивают адекватную "фиксацию закономерностей подъязыка" (Tognini-Bonelli, 2001, p. 55) и позволяют классифицировать составленный подкорпус как представительный для исследования современного состояния украинского юридического дискурса.

Формирование списка коллокаций украинского юридического дискурса

Отсутствие закрытого списка потенциальных коллокаций и корпусный метод их экстракции предусматривает применение в данном исследовании методики открытого списка. Гипотетически в ожидаемый реестр коллокаций должны войти устойчивые сочетания, характерные как для юридического подъязыка, так и украинского языка в целом. Установленные в результате автоматической экстракции модели коллокаций планируется в дальнейшем верифицировать на корпусе большего объема и для изучения текстов других функциональных стилей, в частности научного и публицистического. В этом смысле главной задачей является разработка методики, применимой для описания разных предметных областей (McKeown & Radev, 2000, p. 507). Для автоматического

составления списка в данном исследовании разработана методика, в основу которой положен подход Ф. Смаджа (Smadja & McKeown, 1990), предполагающий выделение коллокаций по статистическим признакам. Статистическая интерпретация основывается на формальном, структурном определении единицы анализа как сочетания двух слов, зафиксированного в исследуемых текстах, по крайней мере, дважды (Marcinkevičienė, 2010, s. 104; Sinclair 1995, p. 57). Использование данного подхода для автоматической обработки подкорпуса позволяет игнорировать условия относительно качественного состава коллокаций, рассматриваемых как привычное, употребительное сочетание двух слов (Sinclair 1995, p. 116; Marcinkevičienė 2010, s. 40; Лендау, 2012, с. 303). Выбор в качестве единицы анализа двусловных контактно расположенных сочетаний объясняется максимальными показателями их частоты в текстах разных функциональных стилей (Перебийніс & Бобкова, 2008, с. 446; Червяк, 2010, с. 44) и наличием разработанного аппарата статистических мер (Хохлова, 2010, с. 16–17). Статистическая интепретация коллокации определяет также объем зональной выборки в 1 млн. слов, организованной из подкорпуса текстов и представленной 129 текстами (не менее 3 тыс. слов каждый) 31 текстового жанра (Табл. 1).

Табл. 1
Диапазон текстовых жанров

№	Жанр	Объем, тыс.
1	Акт	26961
2	Висновок	61438
3	Декрет	29343
4	Договір	79488
5	Закон	3524
6	Інструкція	51883
7	Кодекс	19981
8	Конвенція	30870
9	Концепція	12157
10	Лист	5836
11	Меморандум	15712
12	Методика	52144
13	Наказ	37696
14	Настанова	6678
15	Пакт	8681
16	Перелік	8766
17	Положення	38672
18	Поправки	12657
19	Порядок	57274
20	Постанова	16470
21	Правила	74019
22	Програма	55824
23	Протокол	43630
24	Рекомендації	75520
25	Рішення	25595
26	Роз'яснення	40226
27	Склад	3343
28	Статут	38982
29	Угода	34577
30	Умови	23716
31	Хартія	84731
	ВСЕГО:	1.000.064

В данном случае зоной выборки служит совокупность единиц, установленная в соответствии с порогом частоты употребления сочетания ($f \geq 2$) в текстах объемом в 1,000 064 млн. слов. Таким образом, основными критериями идентификации коллокаций является: 1) наличие двух компонентов устойчивого сочетания; 2) статистический порог частоты их совместного употребления – не меньше двух раз на 1 млн. слов; 3) контактность компонентов; 4) наличие подчинительной связи между компонентами сочетания. Корпусно-ориентированный подход к извлечению коллокаций из организованной выборки текстов реализуеться с помощью процедуры запросов на языке SQL (автор программного обеспечения – В.М. Сорокин). В связи с выбранной в качестве основной модели анализа двухсловной коллокации сочетания с сочинительной связью не принимались во внимание, поскольку данные пары представляют лишь часть большего неоднословного сочетания (*роботодавці та...*). Благодаря возможности задать списком сочинительные союзы, подобные сочетания были автоматически исключены из списка потенциальных коллокаций. Устойчивые сочетания с подчинительной связью рассматриваются с точки зрения формальной структуры как свободные словосочетания (Дарчук, 2013, с. 177–180). При

этом для отображения объективной картины функционирования устойчивых сочетаний в исследуемых текстах учитывался фактический порядок слов в предложении. В результате статистической обработки выборки текстов и лемматизации компонентов пар составлен список потенциальных коллокаций, включающий 64.361 устойчивое сочетание, с указанием частоты употребления (Табл. 2).

Для проверки достоверности статистических данных и определения лексикографически релевантных коллокаций применена асимптотическая гипотеза (Ch. Manning, H. Schütze): $P(w_1w_2) = P(w_1) \times P(w_2)$, где P – вероятность появления слова w в выборке текстов (например, *державна адміністрація* – $P(w_1w_2) = 0{,}003541 \times 0{,}000408 = 1{,}444728e\text{-}06$ при частоте употребления 200 раз на 1 млн. слов).

Табл. 2
База данных потенциальных коллокаций

lemm	lem2	cnt	observatedProb	chast. freq	chast_1. freq	prob1	prob2	probteor
відповідно	до	2361	0,002361	2575	10672	0,002575	0,010672	0,0000274804
договірний	сторона	1902	0,001902	2409	3788	0,002409	0,003788	0,000009125292
у	раз	1406	0,001406	15367	1947	0,015367	0,001947	0,000029919549
конституція	Україна	1301	0,001301	1528	13513	0,001528	0,013513	0,000020647864
згідно	з	1131	0,001131	1506	14119	0,001506	0,014119	0,000021263214
закон	Україна	1032	0,001032	2459	13513	0,002459	0,013513	0,000033228467
міністр	Україна	1006	0,001006	1355	13513	0,001355	0,013513	0,000018310115
під	час	918	0,000918	1438	1618	0,001438	0,001618	0,000002326684
верховний	рада	911	0,000911	1662	1788	0,001662	0,001788	0,000002971656
національний	банк	886	0,000886	1883	3639	0,001883	0,003639	0,000006852237

Полученный список потенциальных коллокаций нуждается в обязательном редактировании в связи с необходимостью: 1) исключения составляющих неоднословных (трех- и четырехсловных) коллокаций (*у, той – у тому числі; міністр, Україна – кабінет міністрів України*), 2) снятия омонимии (*заробітний, платити – заробітний, плата; верховний, рад – верховна рада*) и 3) отбора грамматически правильных коллокаций (*в, вона – в, її*). Результаты исследования тысячи наиболее частотных устойчивых сочетаний, встретившихся в подкорпусе более 59 раз, свидетельствуют, что процент ошибок автоматического распознавания двухсловных коллокаций составляет 15,87 %.

При уменьшении показателя частоты употребления потенциальных коллокаций процент ошибок распознавания увеличивается в основном за счет неправильно идентифицированных двух компонентов трехсловных и четырехсловных сочетаний (*засідання, наглядовий... – засідання, наглядовий, рада; ...справа, Україна – міністерство, внутрішній, справа, Україна*).

Статистический портрет коллокаций украинского юридического дискурса

Использованный в исследовании подход Ф. Смаджа предполагает в первую очередь анализ статистических признаков извлеченных из выборки текстов потенциальных коллокаций. Ранжирование результатов статистической обработки законодательных документах позволяет составить статистический портрет коллокаций украинского юридического дискурса и определить закономерности их употребления в исследуемых текстах (Табл. 3).

Анализ распределения по частоте показывает, что статистическая структура словаря коллокаций украинских законодательных текстов подчиняется действию закона Ципфа: незначительное количество устойчивых сочетаний (менее 1%) употребляется в текстах выборки с максимальной частотой (более 1000 раз), и большая часть извлеченных сочетаний (более 42%) – с минимальной частотой (2 раза). При этом одно устойчивое сочетание – составной

предлог (*відповідно до*) встречается в текстах выборки с максимальной частотой – 2361 раз. Пять устойчивых сочетаний зафиксированы в текстах свыше 1000 раз, среди них: собственно коллокации, лексические сочетания (*договірна сторона* – 1902, *конституція України* – 1301, *закон України* – 1032), характеризующие анализируемую предметную область – юридический дискурс, и коллигации или составные предлоги (*у разі* – 1406, *згідно з* – 1131). Большую часть высокочастотных устойчивых сочетаний составляют лексические сочетания – коллокации (*Верховна Рада, Національний Банк, Президент України, кабінет міністрів*), тем не менее, максимальными показателями частоты характеризуются коллигации – составные предлоги и грамматические коллокации. Однако это наблюдение требует дополнительного исследования в аспекте разработки морфолого-синтаксической классификации коллокаций украинского юридического дискурса и обобщения статистических данных.

Табл. 3
Распределение коллокаций по частоте

№	Количество коллокаций	Частота	%
1	1	2361	0,0015
2	5	1999–1000	0,0075
3	23	999–500	0,036
4	52	499–300	0,081
5	88	299–200	0,137
6	307	199–100	0,477
7	817	99–50	1,269
8	8242	49–10	12,806
9	27218	9–3	42,29
10	27608	2	42,895
	ВСЕГО: 64361		100

Заключение

Основные проблемы извлечения коллокаций из текста составляют установление критериев идентификации, классификация коллокаций и оценка эффективности использованных методов и приемов. Троякая природа коллокаций и несовершенство современных лингвистических процессоров требуют внедрения подхода, объединяющего объективное описание статистических и лингвистических признаков. Предлагаемая методика на основе использования Корпуса украинских текстов позволяет автоматически выделять двусловные сочетания и исчислять их вероятностные характеристики. В основу автоматического формирования списка коллокаций положены формально-статистические критерии идентификации: установление статистического порога употребления и контактность двух элементов устойчивого сочетания, наличие подчитнительной связи. Полученные результаты извлечения потенциальных коллокаций нуждаются в обязательном редактировании с целью исключения составляющих трехсловных и четырехсловных устойчивых сочетаний, снятия лексико-грамматической омонимии и определения грамматически правильных коллокаций. По результатам анализа тысячи наиболее частотных устойчивых сочетаний процент ошибок идентификации составляет 15,87 %. Научную и практическую ценность представляет тестирование описанной методики на материале корпуса украинских текстов большего объема и разных функциональных стилей с целью выявления характерных для украинского языка коллокаций.

References

1. Ananiadou, S. A., 1994. Methodology for Automatic Term Recognition. In: Proceedings of the 15th conference on Computational linguistics, vol. 2, pp. 1034-1038.
2. Bobkova, T., Grydneva, L., Lebedev, K., Kasianenko, M., Lukashevich, V., Petrenko, P., 2009. Corpus of Computational Linguistic Texts. In J. Levická, R. Garabik, ed. NLP, Corpus Linguistics, Corpus Based Grammar Research. Bratislava: Tribun, pp. 35-40.
3. Durrant, Ph., Doherty, A., 2010. Are High-frequency Collocations Psychologically Real? In: Corpus Linguistics and Linguistic Theory, vol. 6, no 2, pp. 125-155.
4. Fontenelle, Th., 1994. What on Earth are Collocations? In English Today: the International Review of the English Language, vol. 10 (40), no. 4, pp. 42-48.
5. Kapstad, M., 2006. Faste uttrykk i russisk og norsk med henblikk på russiskundervisning

for nordmenn: Masteroppgave i russisk språk ved. Våren.

6. Marcinkevičienė, R., 2010. Lietuvių kalbos kolokacijos. Kaunas: Vytauto Didžiojo universiteto leidykla. [Online] available at: http://vddb.laba.lt/fedora/get/LT-eLABa-0001:B.03~2010~ISBN_978-9955-12-656-0/DS.001.0.01.BOOK [Accessed April 2015]
7. McKeown, K. R., Radev, D. R., 2000. Collocations. In: R. Dale, H. Moisl, H. Somers, ed. A Handbook of Natural Language Processing. Marcel Dekker, pp. 507-523. [Online] available at: http://clair.si.umich.edu/~radev/papers/handbook00.pdf [Accessed April 2015]
8. Kotsyba, N., 2013. Praktyczny przewodnik po korpusach języka ukraińskiego. In: Praktyczny przewodnik po korpusach języków słowiańskich. [Online] available at: http://www.domeczek.pl/~natko/papers/przewodnik-korp-ukr2013.pdf [Accessed April 2015]
9. Seljan, S., Gašpar, A., 2012. First Steps in Term and Collocation Extraction from English-Croatian Corpus. In: S. Seljan, ed. Computational Language Analysis, Computer-Assisted Translation and e-Language Learning. Zagreb: Zavod za informacijske studije, pp. 149-156.
10. Seretan, V., 2011. Syntax-Based Collocation Extraction. Berlin: Springer Science & Business Media.
11. Sinclair, J., 1991. Corpus, Concordance, Collocation. Oxford: Oxford University Press.
12. Smadja, F. A., McKeown, K. R., 1990. Automatically Extracting and Representing Collocations for Language Generation. In: Proceedings on the 28-th Annual Meeting of the ACL, Pittsburg: PA, pp. 252-259.
13. Tognini-Bonelli, E., 2000. Corpus Classroom Currency. In: Darbai ir Dienos, no 24, pp. 205-243.
14. Wermter, J., 2008. Collocation and Term Extraction Using Linguistically Enhanced Statistical Methods, Jena: der Friedrich-Schiller-Universit. [Online] available at: http://d-nb.info/993920594/34 [Accessed April 2015]
15. Апресян, Ю. Д. 1995. Избранные труды, том I. Лексическая семантика. Москва: Языки русской культуры.
16. Бобкова, Т., 2015. Лексикографический корпус как источник для словарей нового типа. In A. Diomidova, L. Kamičaitytė, J. Radavičiūtė, ed. Žmogus kalbos erdvėje, no. 8, pp. 31-42.
17. Бобкова, Т. В., 2014. Теоретико-методологічні підходи до вивчення колокацій. In Вісник Київського національного лінгвістичного університету. Серія: Філологія, т. 17, № 2, сс. 14-22.
18. Большакова, Е. И., Клышинский, Э. С., Ландэ, Д. В., 2011. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика. Москва: МИЭМ.
19. Борисова, Е. Г. Захарова, О. В., 1994. Фразеологическое значение в устойчивых словосочетаниях. In Филологические науки, №4, сс. 77-84.
20. Гладка, В. А., 2013. Структурно-синтаксичний підхід у вивченні колокацій (на матеріалі французької мови). In Наукові записки національного університету "Острозька академія". Серія "Філологічна", вип. 39, сс. 16-20.
21. Дарчук, Н., 2013. Комп'ютерне анотування українського тексту: результати і перспективи. Київ: Освіта України.
22. Законодавчі тексти, 2014. [Online] available at: http:// www.mova.info/corpus2.aspx [Accessed April 2015].
23. Зацеркляний, М. М., Узлов, Д. Ю., 2013. Об'єктно-орієнтований тезаурус і словник колокацій для бази знань криміналістичних інформаційних систем. In: Системи обробки інформації, № 2, сс. 183-186.
24. Киверник, Н. Ю. Дифференциальные характеристики коллокаций и коллигаций как несвободных словосочетаний (на примере русских и английских единиц). In Филологические науки, no 4. Синтаксис: структура, семантика, функція. [Online] available at: www.rusnauka.com/33_PRNIT_2012/ Philologia/4_120043. doc.htm [Accessed April 2015]
25. Левицкий, В. В., 2006. Семасиология. Винница: Нова Книга.
26. Лендау, С. І., 2012. Словники: мистецтво та ремесло лексикографії. Київ: К. І. С.
27. Перебийніс, В. І., Бобкова, Т. В., 2008. Частота мовних одиниць як відображення їхніх системних характеристик. In Проблеми загального, германського та слов'янського мовознавства. Чернівці: Книги – XXI, сс. 446-453.
28. Романюк, А., Кваснюк, Г., Романишин, М., 2011. Розпізнавання багатослівних конструкцій. In: Вісник Національного університету "Львівська політехніка". Комп'ютерні системи проектування. Теорія і практика, №711, сс. 158-165.

29. Хайрова, Н. Ф. Узлов, Д. Ю., 2013. Идентификация криминально значимых коллокаций в украиноязычных текстах. In: Збірник наукових праць Військового інституту Київського національного університету імені Т. Шевченка, № 44, сс. 147-151.

30. Хохлова, М. В., 2010. Исследование лексико-синтаксической сочетаемости в русском языке с помощью статистических методов (на базе корпусов текстов): автореф. СПб, 26 с.

31. Червяк, А. В., Вечур, А. В., Шевченко, Е. Л., Ляпота, В. Н., 2011. Об особенностях применения статистических алгоритмов выявления устойчивых словесных цепочек. In: Восточно-Европейский журнал передовых технологий, №4/2(52), сс. 41-47.

32. Шкурко, В. В., 2012. Лексикографічний агент екстракції колокацій у природномовному тексті. In: Вісник Київського національного університету ім. Т. Шевченка. Серія: Літературознавство. Мовознавство. Фольклористика, № 28, сс. 31-35.

33. Ягунова, Е. В., Пивоварова, Л. М., 2011. От коллокаций к конструкциям. In Acta linguistica petropolitana. Труды Института лингвистических исследований. РАН. [Online] available at: http://www.webground.su/data/lit/pivovarova yagunova/Ot kollokatsiy k konstruktsiyam.pdf [Accessed April 2015].

Santrauka

Tatjana Bobkova. Kolokacijų nustatymas ukrainiečių kalbos tekstuose

Straipsnyje aprašoma sudėtinių kolokacijų Ukrainos įstatyminiuose tekstuose atrankos metodika. Dabartinės kolokacijų atrankos metodikos remiasi identifikavimo ir naudojamų procedūrų seka. Darbe analizuojamas tekstyno atrankos metodas ir jo vartojimas, grindžiamas kolokacijų kaip reikšminio statistinio vieneto identifikavimu. Kolokacija suprantama kaip neatsitiktinis dviejų žodžių junginys, reguliariai sutinkamas ir būdingas tam tikriems funkciniams stiliams ir apskritai kalbai. Parengta dviejų žodžių kolokacijų identifikavimo metodika sudaro galimybes, naudojant statistinius metodus ir lemavimo programą, automatiškai atrinkti pastovius dviejų žodžių junginius iš ukrainų kalbos tekstyno. Gauti rezultatai vėliau yra tikslinami dėl galimos homonimijos ir gramatiškai teisingų kolokacijų. Automatinio sąrašo formavimo rezultatų efektyvumo didinimas padės naudotis didesniu kiekiu tekstų ir lingvistiniais kolokacijų identifikavimo filtrais.

About the author

Татьяна Бобкова

Кандидат филологических наук, доцент Киевского национального лингвистического университета, Украина.

Область научных интересов автора

Прикладная лингвистика, корпусная лингвистика, корпусная лексикография, синтаксис.

Адрес

Факультет переводчиков, Киевский национальный лингвистический университет, 73, ул. Велика Васылькивська, Київ, Україна.

E-mail:

tatva93@gmail.com