Висновки. Власне для автоматизації роботи фахівця-лінгвіста пропонується АРМ (автоматичне робоче місце) лінгвіста, особливістю якого є уніфіковане представлення граматичної інформації для трьох мов. З цією метою визначаються частини мови та інші морфологічні характеристики (герундій, артикль, допоміжне дієслово, тощо), що охоплюють перелік можливих характеристик для всіх мов вказаної групи.

1. Оцінка ефективності системи автоматичного морфологічного аналізу (АМА) залежить від обсягів словника, що застосовується, несуперечливості інформації, швидкості опрацювання текстів і можливості аналізу нових слів.

2. Розроблене автоматизоване робоче місце лінгвіста дозволяє максимально оптимізувати розробку аналітичного словника для автоматичного морфологічного аналізу а саме: застосований принцип флективного аналізу на основі позиційно-цифрового кодування допомагає зменшити обсяги словника квазізакінченість, забезпечує компактність збереження лінгвістичних даних, а відтак дозволяє підвищити ефективність системи автоматичного морфологічного аналізу; робить АМА відкритим, оскільки уможливлює аналіз "невідомих для системи" нових слів; способи поповнення бази словоформ – "уведення слова вручну" та "опрацювання повнотекстових масивів" – надає морфологічній моделі змісту, який може відповідати лінгвістичній реальності.

3. Поповнення словника словоформ – як дослідного масиву (базового словника) для формування списку квазіфлексій – завдяки створеному АРМ лінгвіста можна здійснювати двома шляхами: безпосередньо вводячи словоформу вручну або ж за допомогою текстового файлу. Така можливість дозволяє спиратися як на знання мови, скажімо, вводячи унікальні класи словозміни вручну, так і на реальні тексти, які дозволяють реалізувати знання-орієнтовний підхід і до формування морфологічної моделі мови.

4. Єдина система параметризації граматичної інформації дозволяє за допомогою одного АРМ (програмного модуля) опрацьовувати англомовні, російськомовні та україномовні тексти, що є однією з головних вимог до розробки багатомовної системи машинного перекладу.

1. Гельбух А. Ф., Сидоров Г. О. К вопросу об автоматическом морфологическом анализе флективных языков // www.dialog-21.ru/Archive/2005. 2. Лазарева О.Я. Метод формирования словаря квазиокончаний // Інформатизація та нові технології. – № 1. – 1997. – С.9-12. 3. Грязнухіна Т. О., Нікула М. В. Система автоматичного морфологічного аналізу українського наукового тексту // Пробл. україніізації комп'терів. Матеріали 2-ї міжнар. конф. – Київ, 1993. – С. 42-46. 4. Морфологический анализ научного текста на ЭВМ. – К.: Наук. думка, 1989. – 264 с. 5. Замаруєва І.В., Шипнівська О.О. Морфемна обробка текстів в системах машинного перекладу. // Вісн. КНУ ім. Тараса Шевченка. Військово-спеціальні науки. – К., 2008. – №20. – С.61 – 63. 6. Балабін В.В., Замаруєва І.В., Лєнков С.В., Пампуха І.В.Технологічні аспекти реалізації автоматизованих систем машинного перекладу.// Зб. наук. пр. ВІКНУ. – К., 2010. – № 26. – С. 55 – 64.

Надійшла до редколегії 06.06.12

УДК 81'322.2

О.О. Шипнівська, канд. філол. наук

СТВОРЕННЯ БАЗИ ДАНИХ ДЛЯ РОЗРОБКИ АВТОМАТИЧНОГО СИНТАКСИЧНОГО АНАЛІЗАТОРА УКРАЇНОМОВНИХ ТЕКСТІВ (НА МАТЕРІАЛІ ПРОСТОГО РЕЧЕННЯ)

У статті представлено основні засади та специфіку розробки лінгвістичної бази даних для побудови автоматичного синтаксичного аналізатора україномовних текстів на матеріалі простого речення. Головна увага зосереджена на можливостях розроблюваного дослідницького середовища.

Ключові слова: автоматичний синтаксичний аналіз, лінгвістична база даних, просте речення, синтаксична структура речення.

In the paper the instrumental environment for the automatic syntactical analysis of Ukrainian is described. The main characteristics of databases for simple sentence their advantages for the automatic syntactical analysis and general principles of automatic syntactical analysis are presented.

Key words: formal syntactical structure, general principles of automatic syntactical analysis, linguistic databases for automatic syntactical analysis, simple sentence.

1. Автоматичний синтаксичний аналіз: проблеми та перспективи їх вирішення. Проблема розробки алгоритмів автоматичного синтаксичного аналізу є однією з ключових при автоматичному опрацюванні текстів (АОТ) природної мови. Незважаючи на значний доробок у цій царині як теоретичного, так і практичного плану, все ще залишається затребуваним напрацювання відповідних процедур. Актуальною є ця проблема для україномовних текстів у контексті побудови знання-орієнтованої системи машинного перекладу (ЗСМП).

Завдання автоматичного синтаксичного аналізу полягає в побудові формально-синтаксичної структури тексту, представлення речення в синтаксичних категоріях [Апресин 1978; Синтаксический анализ 1999; Дарчук 2005]. У нашій роботі автоматичний синтаксичний аналіз розуміємо як незалежний модуль АОТ текстів природної мови, що покликаний аналізувати та синтезувати всі можливі варіанти синтаксичної структури речення, зокрема україномовного [Замаруева 1999].

Розробка модуля автоматичного синтаксичного аналізу акумулює низку складних практичних та теоретичних проблем [Апресин 1978; Анно 1975; Гладкий 1985; Синтаксический анализ 1985; Баталина 2001; Кобзарева 2007]. Представлені в публікаціях сучасні дослідження проблеми стосовно слов'янських мов зосереджують свою увагу головно на підготовці та побудові синтаксично розмічених корпусів текстів, розробці необхідних для цього засобів аналізу [Кобзарева 2004, Шимкова 2005].

Головним теоретичним підґрунтям автоматичного синтаксичного аналізу, зазвичай, виступає граматика залежностей та граматика безпосередніх складників [Синтаксический анализ 1999]. Представлений у праці підхід, крім цих визначних методів, зорієнтований ще й на врахування комунікативного аспекту синтаксису української мови. Однією з вихідних точок аналізу є поняття правильної синтаксичної структури [Иорданская 1964, с. 215]. Правильною є структура, а у нашій праці, лінійна модель речення, яку може зрозуміти носій мови, не знаючи конкретної семантики слів.

Українській мові як одній із флективних мов притаманна велика кількість типів лінійної структури простих та складних речень. У мовленні можна спостерігати багато структурних компонентів та їх комбінацій. Такий високий рівень синтаксичної розгалуженості ускладнює процеси аналізу, синтезу та моделювання механізмів автоматичного породження тексту. Застосування процедури автоматичного синтаксичного аналізу продемонструвало, що багато синтаксичних одиниць, таких як

складний підмет, складний присудок, однорідні сурядні конструкції, синтаксично неоднозначні структури, потребують розробки спеціальної методики їх попереднього дослідження та моделювання.

Зважаючи на це, опис формальної синтаксичної структури україномовного речення ми пропонуємо здійснювати на основі розроблюваної бази даних як універсального інструментарію, який покликаний унаочнити алгоритми синтезу та аналізу речення, зокрема простого.

2. Загальні принципи автоматичного синтаксичного аналізу в знанняорієнтованій системі машинного перекладу та деякі результати його застосування. Розроблюваний на базі результатів автоматичного морфологічного аналізу та, будучи підготовчим етапом для автоматичного семантичного аналізу, автоматичний синтаксичний аналіз складається з трьох етапів [Толубко 2010]:

- визначення синтаксичних зв'язків між словами у реченні – контекстний аналіз;
- побудова синтаксичної структури речень;
- побудова синтаксичної структури складного текстового цілого.

Контекстний аналіз – це перший крок на шляху до формування розроблюваної бази даних. Здійснюваний повністю автоматично, цей етап визначає синтаксичні відношення між словами та головне слово словосполучення. Ідентифікація синтаксичних відношень реалізується завдяки застосованому словнику синтаксичних правил, який містить контекстуально-синтаксичні правила узгодження, керування та прилягання. Формат умов виконання правил подано в таблиці 1. У першій колонці подано частиномовний код слова, у другій, – завдяки якому слову здійснюється аналіз, третя колонка подає код слова, з яким реалізуються синтаксичні відношення. У четвертій колонці представлено конкретні граматичні значення, завдяки яким здійснюється синтаксичні відношення. П'ята та шоста колонки містять типи синтаксичних відношень та головного слова. У сьомій колонці подано тип здійснюваної операції. Застосований декларативний метод дозволяє лише на основі словника здійснити цей етап аналізу.

На рис. 1 показано приклад речення із застосуванням зазначеного етапу аналізу.

Таблиця 1

Формат представлення правил контекстного аналізу

Частина мови	Завдяки якій ч. мови	З якою ч. мови	Завдяки яким грам. значенням			Тип син такс. відн.	Головне слово	операція
			Рід	Число	Відмін.			
1	2	3	4	5	6	7	8	9
1*		2*	+	+	+	У	1*	М1
1*	24*	1*			+	У	1*/2	М1
…		…	…	…	…	…	…	…
1*		23*			2	К	23*	М1
1	2	3	4	5	6	7	8	9
1*		1*			2	К	1*/2	М2
…		…	…	…	…	…	…	…
14*		9*				П	9*	М3

```
МС =>
        (К1) (ГС)Роль[1*211000000/1*214000000/]органів[1*122000000/]
        (У1) державних[2*922000000/](ГС)органів[1*122000000/]
        (К2) (ГС)в[23*006000000/]системі[1*216000000/]
        (К1) (ГС)системі[1*216000000/]безпеки[1*212000000/]
        (У1) національної[2*212000000/](ГС)безпеки[1*212000000/]
[КР]
```

Рис.1. Приклад застосування правил синтаксичного контексного аналізу україномовного простого речення.

Результати контекстного аналізу опрацьовуються модулем інтерпретації. На цьому етапі аналізу опрацьовуються словосполучення, терміни певної предметної галузі. Для цього кожне слово словосполучення зводиться до початкової форми і зіставляється зі словником семантичних інтерпретацій. Визначені синтаксичні конструкції кваліфікуються як одна лексема і граматична інформація подається до головного слова.

На другому етапі автоматичного синтаксичного аналізу будується синтаксична структура простого речення. Визначаються підмет. Присудок, додаток, обставина, означення.

Застосування цих двох етапів аналізу показало, що існує багато формальних та мовленнєвих синтаксичних структур, які вимагають окремого попереднього аналізу. У багатьох випадках існує потреба розробки окремих модулів, залучаючи різні лінгвістичні дані. Це і стало головним чинником для формування лінгвістичної бази даних для синтаксичного аналізу речення.

3. Розробка бази даних для побудови автоматичного синтаксичного аналізатора україномовних речень: головні принципи та можливості. Ми розглядаємо лінгвістичну базу даних для розробки автоматичного синтаксичного аналізу як дослідницьке середовище. Головне завдання бази:

- представити список моделей лінійної структури речення в українській мові;
- містити всі правильні синтагми, необхідні як для аналізу, так і для синтезу україномовних речень;
- подати всі можливі варіанти синтаксичної неоднозначності та їх частоту в текстах;
- охоплювати всі синтаксичні структури, які потребують спеціальних модулів опрацювання.

База даних розробляється на матеріалі україномовних текстів, зокрема йдеться про прості речення. Формування дослідницького середовища здійснювалось як автоматизовано, так і вручну. На сьогодні база даних складається з 4 таблиць. Перша таблиця містить прості речення, друга – складнопідрядні, третя – складносурядні речення та четверта складні ускладнені речення з кількома типами зв'язку. Вихідними даними для бази слугували результати автоматичного морфологічного аналізу. Для моделювання процесу сприйняття текстової інформації виявили необхідним визначення типу речень. Критерії класифікації подано у таблиці 2.

Таблиця 2

Критерії класифікації речень

Критерії класифікації	Тип речення
Кількість предикативних центрів	Просте
	Складне
	Ускладнене
З/без другорядних членів	Поширені
	Непоширені
	Інше
Тип висловлювання	Розповідне
	Питальне
	Окличне
	Інше

У процесі розробки ми намагалися застосувати універсальний формат представлення лінгвістичної інформації і розглядаємо просте речення як базове стосовно складних речень. Крім типу речення, для кожної такої мовленнєвої одиниці визначались наявність/відсутність предикативного центру, його тип, позиція щодо інших компонентів речення. Формат подання інформації містить таблиця 3.

Таблиця 3

Формат подання лінгвістичного забезпечення

Рівняння	Тип рівняння щодо висловлювання	Наявність предикативного центру	Тип предиката	Тип суб'єкта	Позиція предиката щодо суб'єкта	Модель конструкції з однорідними членами	Тип неоднозначності	Пунктуація

Лінгвістична база даних організована у такий спосіб дозволяє нам отримати необхідну інформацію щодо синтаксичної структури, враховуючи частотні характеристики. З'являється можливість довести або спростувати ті чи інші теоретичні та практичні висновки, отримати нові дані. Скажімо, ми можемо розподілити речення стосовно наявності/відсутності предикативного центру. Наприклад, речення, які містять і підмет, і присудок, становлять 90% досліджуваного матеріалу. Причому з них 53% зафіксовано ускладнені другорядними членами та відокремленими зворотами.

Звісно, при підготовці бази даних головну увагу приділяли кваліфікації членів предикативного центру – підмета та присудка. Важливим залишається встановлення типів членів предикативного центру, моделей їх координації та позиції стосовно один одного. Так, у таблиці 4 подано типи з ймовірністю їх появи у тексті військової тематики.

Таблиця 4

Частотне розподілення типів головних членів речення у текстах військової тематики

Тип присудка	Тип підмета	Відносна частота	Приклад
Простий дієслівний	Простий іменний	71%	дії розпочинаються
Складений іменний	Простий іменний	22%	прикладом є громадянська війна
Складений дієслівний	Простий іменний	5,4%	сторони зобов'язані дотримуватися
Простий дієслівний	Складений іменний	1,4%	207 гелікоптерів будуть оснащені
Складений іменний	Складений іменний		

Дані таблиці є свідченням того, що прості присудок та підмет покривають 56% досліджуваного матеріалу. Речення з предикатом у другій позиції стосовно підмета становить 68% вживаності. Для 32% речень, в яких присудок посідає перше місце щодо підмета можна говорити про його неоднозначність.

Як аналіз, так і синтез україномовного речення потребує породження всіх можливих синтаксичних структур з урахуванням поняття правильної синтаксичної структури. Для виявлення синтаксично неоднозначних конструкцій ми пропонуємо дані нашої бази. Перш за все, слід зауважити, що синтаксична неоднозначність є доволі складною проблемою як теоретичного, так і практичного плану [Колесников 1976]. Необхідність практичного розв'язання задачі виникла лише у контексті проблем машинного опрацювання текстів природної мови, зокрема при машинному перекладі. Так, для російської мови на сьогодні існує чималий доробок із напрацьованим матеріалом. Зокрема, дослідники визначають типи синтаксичної неоднозначності, а вже у контексті автоматичної обробки мови призначають процедури опрацювання таких лінгвістичних фактів [Иорданская 1964; Дрезин 1966]. Проте навіть такі близькоспоріднені мови як українська та російська мають значний потенціал щодо синтаксичного вираження відношень. Наприклад, якщо для російської мови характерний давальний відмінок і об'єкта, і атрибута, то для української мови такі синтагми не є поширеними. Скажімо, в реченні "*Письмо любимой из Парижа*" відношення між першим та другим словом можуть бути визначеними як: об'єктні і як атрибутивні. Аналогічні відношення в українській мові реалізуються засобами родового відмінка, як, наприклад у реченні: "*Тінь яблуні не заважає*".

У нашій праці ми розрізняємо 2 типи синтаксичної неоднозначності: конструкції, в яких суперечливим є визначення головного слова; конструкції, в яких складним для з'ясування є тип синтаксичного зв'язку або ж відношень між членами конструкції. Розроблювана база тільки констатує такі лінгвістичні факти з можливістю їх подальшого опрацювання. Так, у реченні "*Бойовий гелікоптер м-24. Не дочекавшись Росії, Україна почала модернізувати ці повітряні машини із французами"* визначення асоціативного підмета, до складу якого входить елемент "*із французами*" потребує врахування додаткових правил аналізу.

Вимагають розробки окремого модуля аналізу й синтаксичні конструкції з однорідними членами речення, поєднаними сурядним зв'язком. Всі дослідження сурядних конструкцій виходять з принципу, що синтаксична функція одного члена такого ряду дорівнює синтаксичній функції всього ряду в цілому [Санников 1989]. У спеціальній літературі їх формальне представлення кваліфікується однозначно – послідовно від одного компонента до іншого, на практиці такий підхід не є ефективним, а в багатьох випадках істотно спотворює зміст [Падучева 1971; Дарчук 2005]. Стосується це насамперед структур із залежними словами, та тих структур, які виражають різні синтаксичні значення. Так, у реченні "*У спідмінтон можна грати і на даху, і вночі*" за умови зазначеного підходу залишається відкритим питання про кваліфікацію елементів "і на даху, і вночі" як однорідних. Окремо кваліфікуються явища типу: *Підійшло 9 дорослих, молодих самиць*.

Під лінійною структурою речення розуміємо лінійний розподіл важливих синтаксичних одиниць речення – предикативного центра, дієприкметникових та дієприслівникових зворотів, пунктуаційних знаків, які відображають синтаксичну структуру речення. На сьогодні це завдання у базі даних реалізовується тільки для простого речення на матеріалі текстів розмовного жанру (інтерв'ю, тексти телепрограм). Зокрема, найпоширенішою на сьогодні є модель двокомпонентним предикативним центром та відокремленим дієприкметником зворотом $Subj_1$ $Predic_1$ AtrF (де $Subj_1$ – простий іменний підмет, $Predic_1$ – простий дієслівний присудок, AtrF – дієприкметниковий зворот): *Конячка Рябінушка зустрічає дітей біля входу, одразу опиняючись у центрі уваги. Ледь переступивши поріг, малеча поринає в атмосферу свята.*

Перспективи дослідження. На сьогодні база представлена як своєрідний дослідницький інструментарій для наочного робочого уявлення алгоритмів автоматичного аналізу та синтезу україномовних текстів. Матеріал, що міститься в базі, з подальшим розширенням текстової основи може бути застосований не тільки у разі розробки алгоритмів автоматичного синтаксичного аналізу, а й для інших лінгвістичних студій.

1. Анно Е. И. Алгоритм синтаксического анализа предложения / Е. И. Анно // НТИ. – Сер. 2.. – № 7. – 1975. – С. 16-20. 2. Апресян Ю.Д., Богуславский И.М. и др. Лингвистическое обеспечение в системе автоматического перевода третьего поколения / Ю. Д. Апресян, И. М. Богуславский – М., 1978. – 74 с. 3. Баталина А. М., Епифанов М. Е., Кобзарева Т. Ю., Кушнарёва Е. В., Лахути Д. Г. Опыт экспериментальной реализации алгоритмов поверхностно-синтаксического анализа / А. М. Баталина, М. Е. Епифанов, Т. Ю. Кобзарева, Е. В. Кушнарёва, Д. Г. Лахути // Компьютерная лингвистика и интеллектуальные технологии: Труды Междунар. конф. "Диалог'2006" // www.dialog-21.ru/Archive/2006 4. Баталина А.М., Епифанов М.Е., Ивличева О.О., Кобзарева Т.Ю., Лахути Д.Г. Инструментальная среда для экспериментов с алгоритмами поверхностно-синтаксического анализа / А. М. Баталина, М. Е. Епифанов, О. О. Ивличева Т. Ю. Кобзарева, Д. Г. Лахути // Компьютерная лингвистика и интеллектуальные технологии: Труды Междунар. конф. "Диалог'2004" // www.dialog-24.ru/Archive/20064 5. Гладкий А. В. Синтаксические структуры естественного языка в автоматизированных системах общения / А. В. Гладкий. – М.: Наука, 1985. – 140 с. 6. Дарчук Н. П. Дерева залежностей у системах АСА / Н. П. Дарчук // Дарчук Н. П. Комп'ютерна лінгвістика. – К.: ВПЦ "Київський університет". – С. 95-145. 7. Дрейзин Ф. А. Частота появления основных видов синтаксической омонимии в русских текстах / Ф. А. Дрейзин // НТИ. – № 12, 1966. – С. 55-59. 8. Замаруєва І. В. Комп'ютерна модель розуміння природно-мовної текстової інформації / І. В. Замаруєва // Проблемы программирования. –1999. -№2. С.96-102. 9. Иомдин Л. Л., Мельчук И.А., Перцов Н. В. Фрагмент модели русского поверхностного синтаксиса. 1. Предикативные синтагмы // НТИ. – Сер. 2. – № 7. – 1975. – С. 30-43. 10. Иорданская Л. Н. Свойства правильной синтаксической структуры и алгоритм ее обнаружения (на материале русского языка) / Л. Н. Иорданская // Проблемы кібернетики. – Вып. 11. – М.: Наука, – 1964. С. 215-243. 11. Иорданская Л. Н. Синтаксическая омонимия в русаком языке (с точки зрения автоматического анализа и синтеза) / Л. Н. Иорданская // НТИ. – № 5. – 1967. – С. 9-17. 12. Кобзарева Т. Ю. Иерархия задач поверхностно-синтаксического анализа русского предложения / Т. Ю. Кобзарева // НТИ. – Сер. 2. – № 1. – 2007. – С. 23-35. 13. Колесников Н. П. Омонимия в предложении и вопросы ее устранения / Н. П. Колесников – М.: Наука, 1976. – 115 с. 14. Падучева Е. В. О орядке слов в предложении с сочинением: сочинительная проективность / Е. В. Падучева // НТИ. – Сер. 2. – № 3. – 1971. – С. 14-18. 15. анников В. З. Формальное представление сочинительных и сравнительных конструкций / В. З. Санников // Санников В. З. Русские сочинительные конструкции. Семантика. Прагматика. Синтаксис. – М.: Наука, 1989. С. 32-79. 16. Синтаксический анализ научного текста на ЭВМ / Т. А. Грязнухіна, Н. П. Дарчук, В. И. Критская, Н. П. Маловица, Т. К. Пуздырева, К. С. Соломанчук, Л. Г. Братыщенко – К.: Наукова думка, 1999. – 272 с. 17. Толубко В. Б., Шипнівська О. О., Ляшенко А. В. Задачі автоматичної обробки синтаксичної структури в знання-орієнтованій системі машинного перекладу / В. В.Толубко, О. О. Шипнівська, А. В. Ляшенко // Вісник Київського нац. ун-ту ім. Т. Шевченка. Серія: Військово-спеціальні науки. – Вип. 27. – 2010. – С. 136-140. 18. Шимкова М., Гарабик Р. Синтаксическая разметка текстов Словацкого национального корпуса / М. Шимкова, Р. Гарабик // http://korpus.juls.savba.sk

Надійшла до редколегії 25.05.12

УДК 81'374:004.89:[004.738.5:338.46]

В.В. Шкурко

ЛЕКСИКОГРАФІЧНИЙ АГЕНТ ЕКСТРАКЦІЇ КОЛОКАЦІЙ У ПРИРОДНОМОВНОМУ ТЕКСТІ

Статтю присвячено питанням пошуку та ідентифікації колокацій в природномовних текстах. Запропоновано концепцію лексикографічного агента екстракції колокацій. Визначено програмну архітектуру для розподіленого застосування, компонентний склад, функції системи та структури даних. Наведено основні етапи реалізації лексикографічного агента. Обґрунтовано використання запропонованих технологічних рішень.

Ключові слова: автоматизована обробка текстов, колокації, програмні системи.

Article is devoted to the search and identification of collocations in the texts in natural language. Idea of lexicographical agent for collocation extraction proposed. The software architecture for distributed use, components, system functions and data structures are defined. The main steps of the lexicographical agent implementation are given. Use of proposed technological solutions is justified.

Keywords: computer-aided text processing, collocations, software systems

Завдання пошуку та ідентифікації колокацій в природномовних текстах. Лінійна структура природномовного тексту приховує велику кількість елементів різного рівня складності – мовних одиниць, пов'язаних між собою різнотипними зв'язками, які дають свій внесок у загальну семантику тексту. Проблематика автоматичного опрацювання текстів, екстракції знань з низ обумовлює актуальність створення лінгвістичних програм обробки текстових структур, спрямованих на пошук та інтерпретацію зазначених елементів, виявлення та інтерпретацію зв'язків між ними. Серед текстових одиниць, які є об'єктом аналізу, особливу роль відіграють колокації або словосполучення різних типів: фразеологізми, термінологічні словосполучення, еквіваленти слів, усталені словосполучення та ін [1, 2]. У цій статі увагу зосереджено саме на завданні автоматичного пошуку та ідентифікації колокацій. Під колокаціями ми розуміємо синтаксично стійкі й граматично і семантично детерміновані єдності контактно розташованих у тексті слів, між якими існує зв'язок підпорядкування.

Завдання лексикографічного опису мовних одиниць, що складаються з декількох слів, не є новим. Можна