Создание семантического словаря предложных конструкций на основе Украинского национального лингвистического корпуса

Creating a semantic dictionary of prepositional constructions on the basis of the Ukrainian National Linguistic Corpus

Бугаков О. В. (ovbugakov@gmail.com)

Украинский языково-информационный фонд НАН Украины, Киев, Украина

Рассматриваются поисковые возможности Украинского национального лингвистического корпуса, а также создаваемые на его основе лингвистические базы данных. Описана структура электронного семантического словаря предложных конструкций, построенного в соответствии с теорией лексикографических систем.

В последние десятилетия наблюдается тенденция к лексикографированию языковых единиц, формально не являющихся единицами лексического уровня [4, 7]. Попытки лексикографирования семантических, синтаксических, когнитивных и других структур более высокого уровня, чем лексический, не только отражают общую тенденцию к лексикографическому описанию всех языковых явлений, но и отвечают требованиям практики по разработке усовершенствованных систем лингвистического обеспечения [6].

Целью нашего исследования являлось лексикографическое проектирование и создание электронного семантического словаря предложных конструкций в соответствии с принципами теории лексикографических систем. Основой для его создания служил Украинский национальный лингвистический корпус (УНЛК), созданный в Украинском языково-информационном фонде НАН Украины (УЯИФ). Объем корпуса — около 54 млн с/у. Корпус представлен текстами разных стилей и жанров без соблюдения пропорций. В случае необходимости исследователь может самостоятельно создавать подкорпуса отдельных стилей с учетом статистических параметров.

В УНЛК предусмотрены два типа поиска. Первый — по библиографическим реквизитам, второй — полнотекстовый поиск с использованием современных лингвистических технологий. Поиск по библиографическому описанию предназначен, в первую очередь, для отбора подмассива информации для последующей обработки.

Полнотекстовый поиск осуществляется после предварительной процедуры индексирования текстов в кодировке UNICODE, сопоставленных с объектами хранения электронной библиотеки. Для проведения полнотекстового поиска необходимо ввести поисковое словосочетание и задать параметры полнотекстового поиска. Полнотекстовый поиск может быть выполнен с учетом следующих параметров:

- поиск по текстам из текущей корзины;
- с учетом порядка слов;
- с лемматизацией;
- с учетом синонимии;
- по синонимическим рядам;
- по грамматическим параметрам;
- без учета расстояния между словами.

После проведения полнотекстового поиска пользователю предоставляется возможность просмотра локализаций поисковых фраз в выбранном тексте. При выборе одного из объектов результатов поиска происходит поиск контекстов внутри проиндексированного текста.

Поисковые слова контекста в тексте выделяются определенным цветом, например в локализации поисковой фразы *робити добро* красным цветом выделены словоформы *робити* и *добро*, отвечающие поисковой фразе при поиске с лемматизацией (рис. 1).

На материале УНЛК было проведено комплексное исследование функционирования украинских предлогов в украинском тексте на трех уровнях — морфологическом, синтаксическом и семантическом.

Достижение поставленной цели предусматривало решение ряда задач: 1) уточнение реестра предлогов на основе анализа украинских текстов УНЛК; 2) анализ омонимии предлога с другими частями речи в тексте; 3) установление текстовых

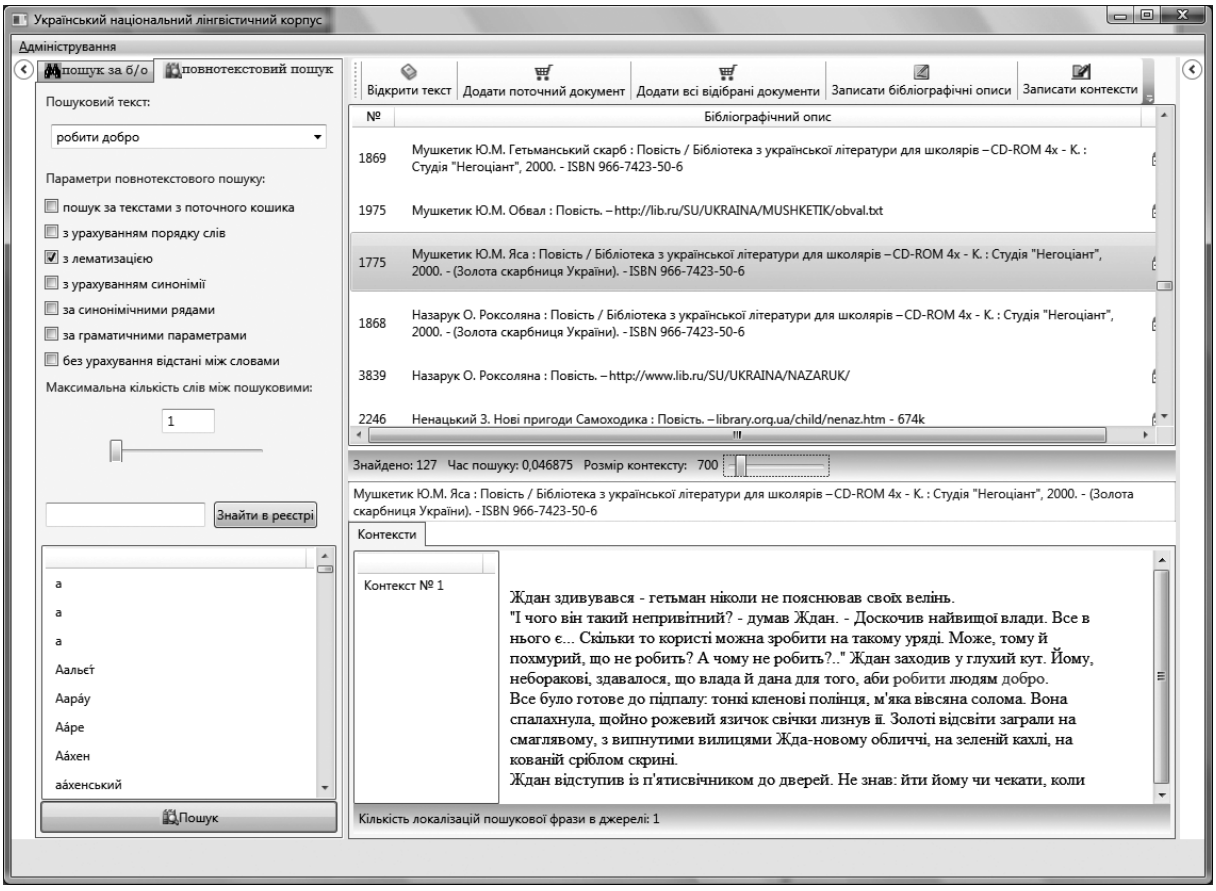


Рис. 1. Полнотекстовый поиск в УНЛК

условий снятия омонимии предлога с другими частями речи в тексте; 4) разработку алгоритма разграничения составных предлогов от сочетаний простого предлога с полнозначным словом (*с помощью, с целью*); 5) разработку алгоритма определения зон предложных связей в тексте как отдельного модуля автоматического синтаксического анализа; 6) установление семантических отношений между компонентами зоны предложных связей в системе автоматического семантического анализа; 7) создание семантического словаря предложных конструкций.

Исследование функционирования предлогов, как и любых других единиц языка, в тексте предусматривает использование статистических методов и метода контекстной диагностики. В связи с этим возникла необходимость проведения анализа на репрезентативном материале с целью обеспечения надлежащего уровня достоверности результатов. Таким материалом и служил УНЛК.

Программное обеспечение УНЛК позволяет создавать специализированные субкорпуса, ориентированные на решение поставленных заданий. С помощью специально разработанной в УЯИФе программы отмеченные субкорпуса переводятся в форматы баз данных с определенной структурой, ориентированной на проведение конкретных лингвистических исследований. Лингвистические базы данных (ЛБД), выполняющие функцию инструмента и материала исследования языкового явления, структурированы по следующему принципу: текстовые сегменты (контексты), которые содержат конкретную языковую единицу (предлог), ставятся в соответствие заранее определенным дифференциальным признакам, по которым осуществляется анализ. Структурирование ЛБД по полям, отвечающим множеству параметров анализа диагностирующих контекстов, и организация доступа к этим полям позволяют автоматически классифицировать материал по каждому из параметров и любой их комбинации.

В соответствии с поставленными выше задачами в УЯИФе создано три предложные ЛБД: лингвистическую базу предложных сочетаний — претендентов на роль составного предлога (ЛБСП), лингвистическую базу грамматических омографов с предложным компонентом (ЛБОП) и лингвистическую базу зон предложных связей (ЛБЗПЗ).

Первая из них (ЛБСП) построена на подкорпусе УНЛК объемом 23 млн с/у. Общая длина ЛБСП — 51025 контекстов [3]. Исходным материалом для второй базы — ЛБОП — служили морфологически размеченные тексты трех стилей (научный, художественный, публицистический), каждый из которых представлен выборкой в 1 млн с/у. Общий объем базы — 200123 контекста [1].

Исходным текстовым материалом для формирования третьей базы — ЛБЗПЗ — служил подкорпус текстов публицистического стиля объемом 6 млн с/у. Общий объем полученной базы — 20768 контекстов [2]. Поскольку именно на основе последней базы и был разработан электронный семантический словарь предложных конструкций, рассмотрим подробнее ее структуру.

База данных была создана для исследования функционирования предлогов на синтаксическом и семантическом уровнях. Она структурирована по полям, соответствующим множеству параметров, предварительно определенных прогнозирующими текстовыми признаками для алгоритмического установления зон предложных связей (ЗПС) при автоматическом семантико-синтаксическом анализе: 1) «Контекст», 2) «Длина ЗПС», 3) «Первая позиция предлога», 4) «Постпозиция ГС», 5) «Контактность ГС», 6) «Главное слово», 7) «Код ГС», 8) «Семантический класс ГС», 9) «Контактность ЗС», 10) «Зависимое слово», 11) «Код ЗС», 12) «Семантический класс ЗС», 13) «Отношение», 14) «Ремарки».

Зона предложных связей включает предлог, главное слово (слово, управляющее предложно-именной синтаксемой) и зависимое слово (слово, подчиняющееся главному с помощью предлога) [2]. На семантическом уровне зоны предложных связей рассматриваются с точки зрения семантической интерпретации синтаксической связи между ГС и ЗС.

При формировании словаря из базы были отброшены строки, в которых: 1) ГС расположено в постпозиции по отношению к предлогу, 2) ГС отсутствует — в случае вхождения предлога в эллиптическую конструкцию, 3) ЗС выражено именем существительным или другой частью речи, являющейся названием книги, газеты, организации и т. д., 4) предлог вошел в состав устойчивого сочетания. В результате использования созданного запроса количество строк в ЛБД сократилось до 13261. Именно столько словарных статей и содержит разработанный словарь.

Теоретическими предпосылками создания словаря является исследование семантики предлога в формализме теории семантических состояний [5], поскольку объектом лексикографирования выступают предложные конструкции, а элементами интерпретационной части — их семантические состояния.

Согласно теории семантических состояний, любое слово (языковая единица) контекста или языкового потока находится в определенном семантическом состоянии, которое для единиц лексического уровня является суммой признаков грамматической и лексической семантики [5]. Семантическое состояние предлога определяем как реализацию конкретного семантического отношения в тексте между главным и зависимым словами, обусловленную семантическими состояниями последних, которые представляют объекты внеязыковой действительности.

Проведение исследования с целью выявления множества типичных семантических состояний класса предлогов является необходимой предпосылкой создания указанного словаря. Установлению семантических состояний предлогов предшествует определение совокупности семантических состояний, которые выражают предлоги с синтаксически связанными с ними словами, с учетом семантических атрибуций ГС и ЗС в ЗПС.

По результатам нашего исследования, проведенного на ЛБД, было выделено 20 типов семантических отношений, которые могут выражать предлоги в тексте:

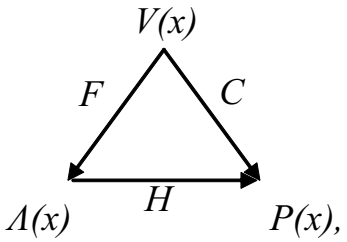
- Объектные (отношение действия к предмету, на который это действие направлено, или предмета к другому предмету, который является объектом действия первого: *за, на, о, замість, на_зміну, між, стосовно, щодо* и др. Этот тип отношений зафиксирован в 16,65% от общего количества ЗПС в ЛБД).
- Пространственные (отношение действия или предмета к месту, пространству, где происходит это действие или находится предмет: *в/у, перед, над, під, біля, поруч_з* и др. — 12,23%).
- Временные (отношение действия ко времени, в которое оно происходит, или явления, предмета ко времени своего существования: *перед, під_час, протягом, після* и др. — 9,29%).
- Условные (отношение действия, признака или предмета к обстоятельствам, условиям, при которых происходит действие или существует предмет: *з_урахуванням, залежно_від, на_випадок, у_випадку, у_разі* и др. — 8,51%).
- Причинные (отношение действия, предмета, явления к причине их возникновения или к их следствию: *унаслідок, у_результаті, у_зв'язку_з, на_ґрунті* и др. — 7,69%).
- Лимитивные (отношение действия к сфере его распространения или предмета к сфере его деятельности: *у галузі, у_межах, у_рамках* и др. — 7,5%).
- Отношения цели (отношение действия к другому действию, явлению, предмету, являющихся целью выполнения этого действия, или к предмету, в интересах которого происходит действие: *в_ім'я, в_інтересах, для, заради* и др. — 7,41%).
- Отношения направления движения (отношение движения в указанном направлении: *до, у_напрямі, усередину, назустріч, мимо* и др. — 5,17%).
- Комитативные (отношение: 1. действия к другому действию, сопровождающему первое, 2. действия к предмету, действием которого сопровождается первое действие, 3. между двумя предметами (лицами), являющихся общими исполнителями или объектами определенного действия: *одночасно_з, паралельно_з, при, спільно_з* и др. — 4,02%).

- Субъектные (отношение действия или предмета к другому предмету, который выполняет действие или является его потенциальным исполнителем: *від_імені, з_боку, за* и др. — 3,67%).
- Отношения способа действия (отношение действия к средству или способу его осуществления: *за_допомогою, за_рахунок, через* и др. — 3,31%).
- Коррелятивные (отношение соответствия действия, состояния, проявления признака к предмету или явлению: *відповідно_до, стосовно_до, згідно_з* и др. — 3,14%).
- Атрибутивные количественные (отношение предмета, свойства, действия к количественному признаку: *біля, близько, понад, коло* и др. — 2,89%).
- Сравнительные (отношение признака, предмета или состояния к другому признаку или предмету, с которыми сравниваются первые: *на_зразок, подібно_до, понад, порівняно_з* и др. — 2,74%).
- Атрибутивные качественные (отношение предмета к его качественному признаку: *з, у_вигляді, з_погляду* и др. — 1,57%).
- Отношения назначения (отношение предмета или явления быть назначением для другого предмета или действия: *для, у_справах* и др. — 1,4%).
- Генеративные (отношение предмета или лица к другому предмету или лицу, указывающих на происхождение первых, или отношения действия к предмету или лицу, являющихся источником этого действия: *від, з, з-під, з-поза* и др. — 1,38%).
- Партитивные (отношение части к целому: *до, в/у, від, з, з-поміж* — 0,89%).
- Функциональные (отношение действия к предмету, на выполнение функций которого направлено действие: *у_ролі, в/у, за* — 0,29%).
- Трансгрессивные (отношение действия, обозначающего преобразование, к предмету, который является результатом или источником этого преобразования: *до, в/у, на, з* — 0,26%).

В тексте каждый из выделенных типов реализует, как правило, определенное множество конкретных отношений в зависимости от семантического состояния конкретного предлога, определяющегося в тексте согласно теории семантических состояний шестью параметрами: релятивной семантикой самого предлога (квазилексическое значение), падежом ЗС, которым управляет предлог (квазиграмматическое значение), а также лексической и грамматической семантикой ГС и ЗС. С учетом указанных параметров в пределах базы было выделено 131 семантическое отношение.

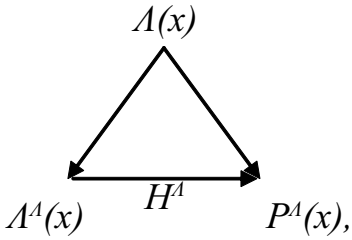
Полученная система семантических отношений предлогов легла в основу электронного семантического словаря предложных конструкций. Интерпретация конкретных предложных сочетаний при создании словаря выводилась путем сопоставления грамматической и семантической информации главного и зависимого слов и потенциальных возможностей предлога передавать семантические отношения.

Словарь представляет собой реализацию определенной лексикографической системы [8]. Структуру словарной статьи $V(x)$ семантического словаря можно представить в виде:



(1)

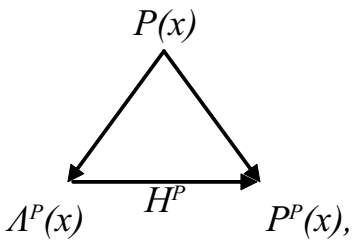
где x — реестровая единица словаря (предложная конструкция); F и C — операторы, выделяющие в тексте формальную и содержательную части описания реестровой единицы x — $\Lambda(x)$ и $P(x)$, соответственно; H — оператор, обеспечивающий соответствие между $\Lambda(x)$ и $P(x)$; элемент $\Lambda(x)$ играет роль левой (реестровой), а $P(x)$ — правой (интерпретационной) частей словарной статьи $V(x)$. Каждая из указанных частей, в свою очередь, делится на левую и правую части, т.е. происходит рекурсивная редукция второго порядка. Структура левой части получает вид:



(2)

где $\Lambda^{\Lambda}(x)$ содержит последовательность символов, а $P^{\Lambda}(x)$ содержит информацию об изъятии из этой последовательности символов трех знаковых репрезентантов компонентов предложной конструкции — k_1 (ГС), k_2 (предлог) и k_3 (ЗС).

Структура правой части имеет вид:



(3)

где в $A^p(x)$ фиксируются грамматические компоненты семантических состояний ГС, предлога и ЗС, а также квазиграмматический компонент семантического состояния предлога, в $P^p(x)$ — лексические и квазилексические компоненты их семантических состояний.

Связи между компонентами семантических состояний ГС, предлога и ЗС можно представить так:

$$\begin{gathered} H(k_1) \\ G(k_1) \rightarrow L(k_1) \\ H(k_2) \\ G(k_2) \rightarrow L(k_2) \\ H(k_3) \\ G(k_3) \rightarrow L(k_3), \end{gathered} \qquad (4)$$

где $G(k_1)$, $G(k_2)$, $G(k_3)$ — грамматические и квазиграмматические компоненты семантических состояний ГС, предлога и ЗС, $L(k_1)$, $L(k_2)$, $L(k_3)$ — лексические и квазилексические компоненты семантических состояний ГС, предлога и ЗС, $H(k_1)$, $H(k_2)$, $H(k_3)$ — операторы, обеспечивающие связи между этими компонентами семантических состояний составляющих предложной конструкции.

В процессе разработки словарной статьи было решено подавать грамматические и квазиграмматические компоненты семантических состояний ГС, предлога и ЗС в левой части словарной статьи, что обусловлено принципом экономичности.

В соответствии с этим схему словарной статьи словаря можно представить следующим образом:

ГС *ч.р.* **ПРЕДЛОГ ЗС** *ч.р. в …пад.* || семантический класс ГС; семантический класс ЗС; семантическое отношение.

ГС подается в начальной форме, а ЗС — в форме того падежа, которого требует конкретный предлог. Компонент, отвечающий грамматическому значению предлога, а именно, падеж зависимой именной формы, подается не возле предлога, а в блоке грамматического значения ЗС. Для увеличения визуального эффекта левая часть отделена от правой двумя прямыми черточками. Вместо семантических отношений подаются только их названия, а сами толкования поданы в отдельном файле. Такая форма представления будет упрощать автоматический поиск предложных конструкций по параметру семантического отношения.

Проиллюстрируем изложенное на примере словарной статьи:

ПРИХОДИТИ *дієсл.* **НА БАЗАР** *ім. у знах. в.* || рух; місце; Відношення напрямку руху 1.

Структурогенными компонентами левой части являются:
k_1 — **ПРИХОДИТИ** (ГС), k_2 — **НА** (предлог), k_3 — **БАЗАР** (ЗС).

Структурогенными компонентами правой части являются:

$G(k_1)$ — «глагол» (грамматический компонент семантического состояния ГС), $G(k_2)$ — *знах. в.* — «винительный падеж» (квазиграмматический компонент семантического состояния предлога), $G(k_3)$ — *ім.* — «имя существительное» (грамматический компонент семантического состояния ЗС).

$L(k_1)$ — рух (лексический компонент семантического состояния ГС), $L(k_3)$ — місце (лексический компонент семантического состояния ЗС), $L(k_2)$ — Відношення напрямку руху 1 (квазилексический компонент семантического состояния предлога).

Структурирование словарной статьи разработанного нами словаря предусматривает возможность поиска информации по всем выделенным структурогенным компонентам: по знаковому представлению предлога, ГС и ЗС, по их грамматическим показателям, то есть по частеречной принадлежности, по падежу ЗС, по семантическим классам ГС и ЗС и по семантическим отношениям. Поиск возможен как по отдельным параметрам, так и по их совокупности.

Словарь представляет собой открытую систему, которая может постоянно пополняться новыми предложными конструкциями.

Понятно, что возникает вопрос в целесообразности создания такого типа словаря. Кроме того, что словарь можно использовать в дальнейших исследованиях, в частности для изучения синонимии и антонимии предлогов, словарь может быть использован в системах автоматической обработки текста, в частности в лингвистических анализаторах как источник лексической информации при идентификации ЗПС в тексте. Рассмотрим возможность использования словаря в синтаксическом анализаторе. На этапе синтаксического анализа обращение к словарю происходит в случае, когда на основе грамматических, позиционных и семантических признаков не удается однозначно установить главное слово в ЗПС из-за присутствия нескольких формальных претендентов на роль ГС. Например, в предложении:

«Але<SC> ***проходження<NA>*** *програми<FB> дій<FI> уряду<MB>* ***через<PD>*** ***парламент<MD>*** *показало<VT>, що<SS> про<PD> взаєморозуміння<ND> та<SC> взаємопідтримку<FD> не<Z0> йдеться<YO>.<e>»*

у предлога *через* есть несколько претендентов на роль ГС — имена существительные в препозиции *проходження, програм, дій, уряду* и глагол-сказуемое в постпозиции *показало*, которые входят в пределы интервала поиска ГС. Идентификация претендента *проходження* с конструкцией в словаре указывает на то, что именно эта словоформа является главным словом в зоне связей предлога *через*.

Литература

1. *Бугаков О. В.* Аналіз граматичної омонімії прийменників у мові й у тексті // Мовознавство. К.: 2004, № 5–6, С. 87–98.
2. *Бугаков О. В.* Зони прийменникових зв'язків у синтаксичній структурі українського речення // Мовознавство. К.: 2005, № 5, С. 75–87.
3. *Бугаков О. В., Грязнухина Т. А., Рабулец А. Г.* Формирование предложных текстоориентированных баз данных на корпусе украинских текстов // Труды международной конференции «MegaLing'2005. Прикладная лингвистика в поиске новых путей». СПб.: Изд-во «Осипов», 2005. С. 11–16.
4. *Золотова Г. А.* Синтаксический словарь: Репертуар элементарных единиц русского синтаксиса. М.: Наука, 1988.
5. *Корпусна лінгвістика / В. А. Широков, О. В. Бугаков, Т. О. Грязнухіна* та ін. К.: Довіра, 2005.
6. *Рабулець О. Г., Сухарина Н. М., Широков В. А., Якименко К. М.* Дієслово в лексикографічній системі. К.: Довіра, 2004 .
7. *Русский* семантический словарь. Толковый словарь, систематизированный по классам слов и значений / Под ред. Н. Ю. Шведовой. М., 1998. Т. 1.
8. *Широков В. А.* Інформаційна теорія лексикографічних систем. К.: Довіра, 1998.