О.В. Бугаков

ИСПОЛЬЗОВАНИЕ УНЛК В ЛИНГВИСТИЧЕСКИХ ИССЛЕДОВАНИЯХ

1. Описание УНЛК

Статья посвящена обзору некоторых исследований, проведенных на украинском национальном лингвистическом корпусе (УНЛК), созданном 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 предыдущей 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:

- с учетом порядка слов;
- с лемматизацией;
- с учетом синонимии;
- по синонимическим рядам;
- по грамматическим параметрам;

– без учета расстояния между словами;
– семантический поиск.

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 *робіть* и *добро*, что отвечают поисковой фразе при поиске с лемматизацией (рис. 1).

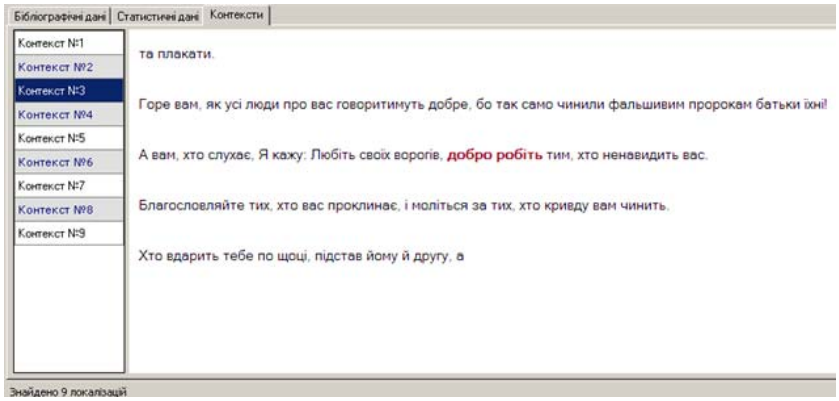

Рис. 1. Просмотр контекстов

Каждое слово, входящее в контекст, является активным. При вызове контекстного меню (нажатие правой клавиши мыши на слове) пользователю открывается возможность получения развернутой справочной информации из грамматического или толкового словаря, или представляется информация на базе словаря синонимов в виде развернутого списка синонимических единиц, сгруппированных по синонимическим рядам (рис. 2). Такая функ-

ция возможна лишь для слов, процедура лемматизации для которых прошла успешно.

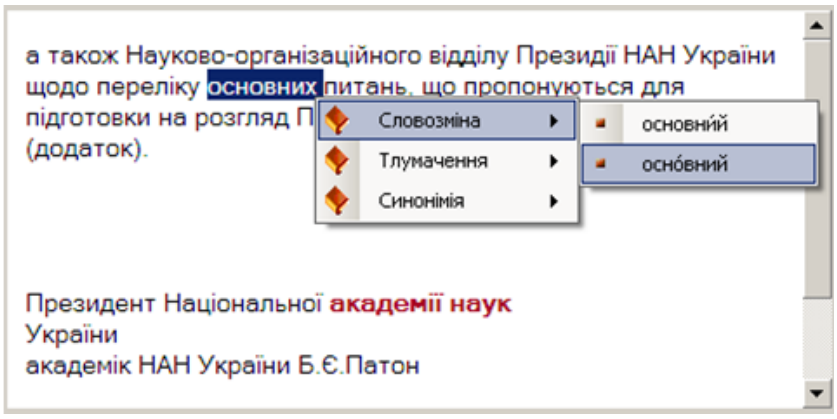

Рис. 2. Связь контекстов со словарями

2. Исследования, проведенные с помощью УНЛК

Ниже приведен список некоторых исследований, в которых был использован УНЛК:

2.1. Исследование грамматической омонимии

Для анализа функционирования морфологических омонимов в текстах современного украинского языка в пределах УНЛК были сформированы три подкорпуса научного, художественного и публицистического стилей, каждый из которых содержал по 1 млн с/у. В результате анализа полученных подкорпусов были получены статистические данные по соотношению в них однозначных и омонимичных единиц, в т.ч. отдельно по указанным стилям. Были получены статистические данные по соотношению межчастеречных омонимов к общему количеству морфологических омонимов, а также данные по каждому типу морфологической омонимии. Информация о реализованных в текстах типах

межчастеречных омонимов и частоте их употребления является важной для разработки алгоритмов автоматического снятия морфологической омонимии. Выделенные в текстах типы морфологической омонимии обуславливают выбор конкретных текстовых ситуаций, а по данным о частоте моделей определяется иерархия правил алгоритма[1].

2.2. Исследование синонимичности текстов

В связи с распространением т. наз. студенческого плагиата возникла необходимость исследования этого явления и определения близких текстов на основе понятия слабой синонимичности текстов. Критерий синонимичности текстов: если больше *n*% слов текста *T*1 присутствуют в тексте *T*2 в тех же формах, если больше *m*% пар слов, которые стоят рядом в тексте *T*1, присутствуют в тексте *T*2 в тех же формах, причем слова должны стоять не обязательно рядом, а на расстоянии не больше l, и *k*% троек слов текстов *T*1 и *T*2 совпадают, причем слова должны стоять на расстоянии не больше l, то тексты являются слабо синонимическими. Числа *m*, *n*, *k* зависят от длины исследуемых текстов[2].

2.3. Проведение лингвистических экспертиз

Одним из подкорпусов УНЛК является корпус законодательства Украины. На основе этого корпуса было проведено несколько лингвистических экспертиз. Первая касалась программ украинских политических партий и блоков. В результате были автоматически построены частотные конкордансы текстов программ 12 политических партий (и, соответственно, блоков), а также частотный конкорданс текста Конституции Украины[3]. Вторая экспертиза касалась проекта нового Налогового Кодекса Украины.

---

[1] *Корпусна* лінгвістика / В.А. Широков, О.В. Бугаков, Т.О. Грязнухіна та ін. К.: Довіра, 2005. 471 с.

[2] Там же.

[3] Там же.

Результаты проведенной экспертизы были переданы в Кабинет Министров Украины для внесения соответствующих корректив в проект кодекса.

Кроме того, была семантически размечена Конституция Украины, т.е. каждому слову приписывалось именно то лексическое значение или оттенок, в котором оно употреблялось в контексте. Это стало возможным благодаря интеграции толкового словаря в корпус. В результате такой разметки выяснилось, что в толковом словаре отсутствуют некоторые значения и оттенки, которые присутствуют в Конституции.

2.4. Создание толкового словаря и словаря синонимов

В УЯИФе создается 20-томный толковый словарь украинского языка, а также его электронная версия. УНЛК является основным источником иллюстраций значений слов. Широкие поисковые возможности помогают подобрать наиболее точные иллюстрации. Мы можем искать слово «как есть», т.е. именно в той грамматической форме, какая нам необходима; искать слово/словосочетание во всех грамматических формах; искать слово/словосочетание только в указанных грамматических формах; а также воспользоваться поиском по маске (поиск по какой-нибудь части слова).

УНЛК также используется как материал при создании словаря синонимов. Составители словаря имеют возможность смотреть контексты, в которых встречаются те или иные слова, входящие в синонимический ряд.

2.5. Исследование функционирования украинских предлогов в тексте

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

Достижение поставленной цели предусматривает 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 условий снятия омонимии предлога с другими частями речи в тексте; 4) разработка алгоритма разграничения составных предлогов от сочетаний простого предлога с полнозначным словом (*с помощью, с целью*); 5) разработка 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. Структуризация 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 – 51 025 контекстов[1]. 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 – 200 123 контекста[2]. Третья база – ЛБЗПЗ – сформирована на основе подкорпуса текстов публицистического стиля объемом 6 млн с/у. Общий объем полученной базы – 20 768 контекстов[3]. На основе последней базы был разработан электронный семантический словарь предложных конструкций.

---

[1] *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. // URL: **http://ovbugakov.org.ua/art_site/article6.htm**

[2] *Бугаков О.В.* Аналіз граматичної омонімії прийменників у мові й у тексті // Мовознавство. 2004. № 5–6. С. 87–98. // URL: **http://ovbugakov.org.ua/art_site/article2.htm**

[3] *Бугаков О.В.* Зони прийменникових зв'язків у синтаксичній структурі українського речення // Мовознавство. 2005. № 5. С. 75–87. // URL: **http://ovbugakov.org.ua/art_site/article7.htm**)

Помимо поставленных и решенных задач данного исследования, полученные базы данных сами быть инструментом других исследований, на основе которых автоматически будут формироваться новые базы данных.

В статье описаны лишь некоторые исследования, которые проводились с использованием УНЛК. Корпус может служить для создания новых баз данных, ориентированных на конкретные лингвистические исследования, для получения различных статистических данных, для проведения различных лингвистических экспертиз, для создания новых словарей, а также для решения проблем синтаксической и семантической разметки и развития поисковых технологий.