

ПРИМЕНЕНИЕ ЯЗЫКА ЗАПРОСОВ ЯНДЕКС ДЛЯ ПОИСКА БЕЛОРУССКОЯЗЫЧНЫХ КОНТЕКСТОВ К ЗАДАННОМУ ЛСВ СЛОВА

Обращение к массиву текстов Интернета как к источнику лексико-графического материала и (реже) как к инструменту верификации лингвистических гипотез достаточно традиционно. Для русскоязычного сегмента Интернета информационно-поисковой системой, наиболее успешно выполняющей функции лингвистического поиска по коллекции текстов сети, изначально был Яндекс (см. анализ интернет-ИПС как инструмента лингвистических исследований в датирующейся 2005 годом работе В. П. Захарова [1] и примеры использования Яндекса как инструмента лексикографа в вышедшей годом ранее работе В. И. Беликова [2]). Заметим, что применительно к вышеозначенной задаче у Яндекса до сих пор нет конкурентов среди ИПС.

С развитием корпусов как таковых (и корпусов русского языка — в частности) поиск языковых данных по модели “ИПС как корпус-менеджер плюс текстовые ресурсы Интернета как корпус” утратил свою актуальность, но идея такого поиска претерпела изменения и легла в основу концепции создания сверхбольших корпусов на основе веб-текстов (Web as Corpus); описание проблем, связанных с несбалансированностью состава и сложностью определения системы разметки ресурсов, созданных в рамках этой концепции, дано в работе В.П. Захарова [3].

Применительно к задаче поиска белорусскоязычных контекстов инструментом первой очереди является Беларускі N-корпус (далее — БН) [4]; сравнительная характеристика этого и других свободно доступных корпусов белорусского языка дана в [5]. Несмотря на представительный объем вышеназванного ресурса, в некоторых случаях контексты к заданному ЛСВ отсутствуют либо по той или иной причине отклоняются лексикографом. Одним из альтернативных инструментов поиска белорусскоязычных контекстов является Яндекс, обладающий гибким языком запросов и ведущий поиск с учетом не только русской, но и белорусской морфологии [6].

Приведем примеры решения нескольких лингвистически-ориентированных поисковых задач.

Задача-1. Найти контексты для слова *фортка*, обозначающего дефект учебного расписания (свободный от занятий промежуток между парами в пределах одного дня).

Решение: [domain:by фортка & (расклад | універсітэт | заняткі | дэканат)].

Здесь и далее текст запроса набираем в квадратных скобках.

Комментарий. В запросе использован прием определения лексического окружения. Поиск ведется по доменной зоне **by**; оператор **&** определяет поиск контекстов, где словоформа слова *фортка* находится в одном предложении со словоформой какого-либо слова из списка (**расклад | універсітэт | заняткі | дэканат**). Среди первых 10 документов выдачи Яндекса релевантные контексты содержатся в 9.

Пример контекста: *Мне склалі такі нязручны **расклад**, столькі «**фортак**», што я ў школе буду знаходзіцца з раніцы да вечара* (<http://budzma.by>).

В БН поиск с параметрами «Усе словаформы» и «Любая адлегласць у межах абзаца» пар *фортка — расклад, фортка — універсітэт, фортка — заняткі, фортка — дэканат* дает 1 релевантный контекст: *Правяраць сшыткі і пісаць планы нельга непасрэдна на ўроку, а вось забараніць рабіць гэта падчас перапынкаў паміж **заняткамі** ніхто не можа. <...> «**Форткі**» ніяк не аплачваюцца і на сваёй сутнасці гэта — асабісты час настаўніка, які можа падчас паўзы пачытаць кнігу, пагуляць на парку ці нават схадзіць у краму* (БН: Звязда).

Задача-2. Найти контексты для слова *аб'яднанне* в ЛСВ-1 'дзеянне паводле значэння дзеяслова аб'яднаць' и в ЛСВ-2 'вайсковае фарміраванне'.

Решение для ЛСВ-1: [domain:by аб'яднанне /+1 (з | у)].

Комментарий. В искомом ЛСВ слово имеет обязательную валентность, эксплицируемую, помимо окончаний, и предлогами *з, у (ў)*, поэтому в запросе может быть использован прием определения грамматического окружения. Фрагмент запроса «/+1» определяет поиск контекстов с предлогами *з* или *у (ў)* непосредственно справа от любой словоформы слова *аб'яднанне*.

Заметим, что для решения задачи-1 запрос, определяющий грамматическое окружение (а именно: [domain:by фортка /+1 у]) будет неуспешен: эту же факультативную валентность определяет слово *фортка* в ЛСВ-2 'частка окна' и ЛСВ-3 'брамка'.

Решения для ЛСВ-2.

[(domain:by | domain:org) аб'яднанне /(+1 +2) (!войск | !войскаў | !армій)]

Комментарий. Этот вариант запроса определяет лексико-грамматическое окружение. Восклицательный знак около элементов группы (!войск | !войскаў | !армій) обозначает, что будет осуществляться поиск контекста именно с этой словоформой.

Фрагмент запроса «/(+1+2)» означает, что два текстовых фрагмента, определенных запросом, следуют либо непосредственно друг за другом, либо между ними находится 1 слово. В результате будут найдены контексты со следующими синтагмами: *аб'яднанне ±{аўтаномных} армій, аб'яднанне ±{авіяцый і, ВПС і, белагвардзейскіх, казацкіх, паветраных, расійскіх, рускіх, савецкіх, саюзных, сухапутных, чыгуначных} войск(аў)*.

Фрагмент запроса (domain:by | domain:org) определяет список доменных зон для поиска контекстов.

Пример контекста: *Фронт (Ф) — фарміраванне, вышэйшае, да Другой сусветнай вайны, апэратыўна-стратэгічнае аб'яднанне войскаў (сіл) узброеных сіл на кантынентальным тэатры ваенных дзеянняў, прызначанае для правядзення стратэгічных ваенных аперацый па знішчэнні буйных варожых групавак або абароны жыццёва важных тэрыторый* (be.wikipedia.org).

Приведем другой вариант решения, где задается лексико-грамматическое определение левого окружения слова в искомом ЛСВ:

[(domain:by | domain:org) (камандуючы | камандаваць) /(+1 +3) (!аб'яднаннем | !аб'яднаннямі)]

Пример контекста: *У чэрвені 1940 года войскі акругі ўзначаліў генерал-палкоўнік Д.Р. Паўлаў, якому да гэтага не даводзілася камандаваць аб'яднаннямі: корпусам, арміяй, буйнымі групоўкамі войск у значных аперацыях, ён не прайшоў адпаведнай штабной школы* (<http://www.nashkraj.by>).

Также возможен вариант решения с определением лексического окружения:

[(domain:by | domain:org) аб'яднанне & (войска | камандуючы | камандаваць | армія)]

Комментарий. Оператор & задает поиск в пределах предложения словоформ слова *аб'яднанне* і словоформ любого слова списка. В случаях, когда запрос подобной структуры слишком узок и не дает контекстов, зону поиска можно расширить следующими способами:

а) заменить оператор & на фрагмент «/(1 12)»; компоненты, объединенные таким оператором, будут искаться на расстоянии от 1 до 12 слов вправо / влево, то есть могут быть найдены в разных предложениях;

б) заменить оператор & на оператор &&; компоненты, объединенные этим оператором будут искаться в пределах документа.

Задача-3. Найти контексты для слова *адказнік*, где реализованы ЛСВ, отличные от ЛСВ 'адна з бакоў грамадзянскай справы ў судзе'.

Решение.

[(domain:by | domain:org) адказнік ~ (суд | паседжанне | працэс | адвакат | прадстаўнік)]

Комментарий. Оператор ~ исключает из выдачи страницы со словоформами слов списка (суд | паседжанне | працэс | адвакат | прадстаўнік). Оставшиеся в выдаче документы содержат преимущественно страницы с примерами реализаций ЛСВ по условию задачи.

Задача-4. Найти контексты для слова *кавярня* в русскоязычном и белорусскоязычном сегментах байнета.

Решение: поиск белорусскоязычных контекстов:

[domain:by lang:be кавярня].

Комментарий. Код языка и код доменной зоны различны (**be vs by**)!

Поиск русскоязычных контекстов:

[domain:by lang:ru кавярня].

Комментарий. По этому запросу будут найдены также страницы с белорусскоязычными вкраплениями-названиями (*Мая кавярня* и т.п.).

Задача-5. Найти контексты с косвенными формами заданного слова.

Решение. Формат запроса таков: [слово -!слово].

Запрос такого формата пригоден также для проверки наличия описания целевого слова в лингвистической базе данных Яндекса. Пустая выдача по запросу такого формата (например: [альтэрнатыўнік -!альтэрнатыўнік]) свидетельствует о том, что слово не разобрано Яндексом, и для представления слова в полном наборе словоформ необходимо в запросе перечислить все словоформы этого слова внутри круглых скобок через вертикальный слеш.

В общем случае универсальным приемом при поиске контекстов для заданного ЛСВ является прием определения лексического окружения; его дополняет прием определения грамматического окружения, успешный, если слово в искомом ЛСВ имеет обязательную и эксплицируемую раздельнооформленной единицей валентность. В завершение сошлемся на конспективное изложение рекомендаций И. В. Сегаловича по лингвистическому поиску средствами ИПС [7] и на работу В. И. Беликова, В. П. Селегея и С.А. Шарова [8], в которой также рассмотрены проблемы лингвистического поиска средствами корпусов и традиционных ИПС.

Подготовлено в рамках проекта, реализуемого при поддержке БРФФИ (договор № Г15Р-029 от 04.05.2015).

ЛИТЕРАТУРА

1. Захаров, В. П. Веб-пространство как языковой корпус [Электронный ресурс]. — Режим доступа: <http://www.dialog-21.ru/archive/2005/zakharov%20v/zakharovv.htm>. — Дата доступа: 07.03.2016.
2. Беликов, В. И. Яндекс как лексикографический инструмент [Электронный ресурс]. — Режим доступа: <http://www.dialog-21.ru/archive/2004/belikov.htm>. — Дата доступа: 07.03.2016.
3. Захаров, В.П. Оценка качества Интернет-корпусов русского языка // Корпусная лингвистика—2015: сб. трудов междунар. науч.конф. — СПб., 2015. — С. 219–229.
4. Беларускі N-корпус [Электронный ресурс]. — Режим доступа: <http://bnkorporus.info/>. — Дата доступа: 05.11.2015.
5. Станкевич, А. Ю. Программные инструменты создания конкордансов для белорусскоязычных опорных единиц // Вестник МГЛУ. Сер. 1. Филология. — 2016. — № 1. — № 1 (80) — С. 108–115.
6. Язык запросов Яндекс [Электронный ресурс] / Яндекс. — Режим доступа: <https://yandex.ru/support/search/query-language/qlanguage.xml>. — Дата доступа: 07.03.2016.
7. Сегалович, И. Проблемы в использовании поисковых систем Интернета как лингвистического ресурса [Электронный ресурс] // Оптимизация. — Вып. 148. — Режим доступа: <http://www.optimization.ru/subscribe/148.html>. — Дата доступа: 07.03.2016.
8. Беликов, В. И., Селегей, В.П., Шаров, С.А. Прологомены к проекту Генерального интернет-корпуса русского языка [Электронный ресурс]. — Режим доступа:

<http://www.dialog-21.ru/digests/dialog2012/materials/pdf/162.pdf>. — Дата доступа:
07.03.2016.

