Национальный корпус русского языка и проблемы гуманитарного образования. Материалы международной научной конференции. Москва 19-20 апреля 2007. М. 2007. С. 44-46.

Специфика поиска в Национальном корпусе русского языка: что важно знать о ней преподавателю-словеснику

Крылов С. А.
Институт востоковедения РАН
krylov-58@mail.ru

Поиск информации в Национальном корпусе русского языка (далее НКРЯ) характеризуется некоторыми особенностями, которые важно учитывать преподавателям-словесникам, использующим поиск по НКРЯ в своей профессиональной практике.

Поиск бывает двух типов: «поиск точных форм» (далее ПТФ) и «лексико-грамматический поиск» (далее ЛГП). ПТФ русского языка (далее РЯ) аналогичен поиску в Яндексе и потому не нуждается в пояснениях. Между тем ЛГП учитывает грамматические и семантические признаки слов. Есть два вида ЛГП – по подкорпусу со снятой (далее ЛГП-1) и с неснятой (далее ЛГП-0) омонимией.

Проиллюстрируем специфику ЛГП-0 в НКРЯ на конкретном примере. Допустим, что в поисковом запросе – «существительные в им. п. со значением сверхъестественного существа (‘superanim’) + глагол в изъявительном наклонении со значением движения».

Среди первых 102 реакций на этот запрос оказалось 22 предсказуемых исходя из фактов, не связанных со спецификой ЛГП-0 в НКРЯ. Из них 16 адекватных запросу (*титаны ушли, Ангел приходит, Призрак покидает, чудовища выходят, Бог приходит, Ангел пролетел, ангелы летают, Призрак бродит, СНЕГУРОЧКИ ИДУТ, Снегурочка поехала, лешие бродят, чудовище сбегает, привидения гуляют , ведьмы летают , черт пролезет , ЛЕ́ШИЙ БРО́ДИТ*); 2 неадекватны в силу лексической и морфологической омонимии (при ЛГП-0) (*сирены провёл, Чебурашки идёт*); 1 правильна, но синтаксически не согласована в силу однородности подлежащего (*Кощей выходят*); и 1 псевдоправильна в силу контактной связи со сравнительным подлежащим (*зомби бродил*). Остальные 80 реакций объясняются спецификой ЛГП-0 в НКРЯ.

1. При ЛГП учитывается лишь «лексико-морфологический» состав сегментов, а их пунктуация не принимается во внимание. Поэтому два сегмента, совпадающих по «лексико-морфологическому» составу, но отличающиеся расстановкой знаков препинания (например, сегмент «*он идёт*» и сегмент «*он, идёт*»), в ЛГП будут иметь одинаковые поисковые образы. Среди 102 реакций 24 неадекватных, где неадекватность объясняется наличием пунктуационных знаков между членами сочетания (из них 15 сочетаний с запятой, 7 с кавычкой, 1 с дефисом, 1 с «точкой с запятой»).

2. Если ограничиваться лишь “документированным” ЛГП-0, то у

неискушённого пользователя может возникнуть впечатление, что импликационные отношения между признаками одного сегмента не принимаются во внимание. Так, при запросе – «N в им. п. со значением сверхъестественного существа (‘superanim’)», то среди ответов будут такие, как *господа* (5), *гора* (2), *сатира* (1), *черта* (1), *горе* (1), *Кощеев* (1).

В РЯ все конкретные экземпляры абстрактных сегментов типа *черта*, *господа*, *гора*, *сатира*, *горе*, (ср. *бук*, *Драконов* и т. п.) обладают таким свойством: ЕСЛИ сегмент обозначает ‘superanim’, ТО падеж словоформы не является “nom.”; и наоборот, ЕСЛИ он имеет признак “nom.”, ТО он не обозначает ‘superanim’. Если в запросе признаки “nom.” и ‘superanim’ сочетались КОНЪЮНКТИВНО (а по здравому смыслу это именно так), то «на выдаче» пользователь получает в ответ набор употреблений не только таких сегментов, где признаки «им. п.» и ‘superanim’ связаны в РЯ тоже конъюнктивно, но и набор употреблений таких словесных сегментов, у которых признаки «им. п.» и ‘superanim’ в РЯ связаны отношением взаимоисключения (т. е. в любом из употреблений такого сегмента представлен лишь 1 из этих признаков, а не оба сразу) и т. о. связаны не отношением совмещения, а отношением взаимоисключения. Так же соотносятся признаки «N» и ‘superanim’ у сегментов типа *Драконов*, *Кощеев*.

3. Следует считаться с тем, что лексико-морфологические гипотезы, используемые в ходе ЛГП, нередко выходят за рамки норм РЯ. Так, признак ‘superanim’ усматривается у слова *гор*, поэтому при ЛГП-0 примеры с сегментами *гор* (2 примера) и *горы* (4 примера) фигурируют среди реакций на вышеуказанный запрос. Если усомниться в существовании лексемы *гор* с признаком ‘supeanim’ или в её принадлежности к норме РЯ, то другие сегменты с той же основой также допускают отнесение к этому подклассу.

4. Учителю-словеснику, прибегающему к ЛГП, следует принять во внимание, что наименования семантических признаков, принятые в системе ЛГП, далеко не всегда прямо соответствуют одноименным терминам для обозначения семантических признаков, используемым в современных толковых словарях, грамматиках или исследованиях по семантике. Между тем при описании семантических признаков в НКРЯ не указаны научные источники, послужившие основой для приписывания семантических признаков. В результате мотивы принятия тех или иных решений остаются как бы «за кадром». Так, неясен мотив приписывания признака ‘superanim’ сегменту *тварь* (2 раза) или признака «движение» сегментам *задрали* (1 раз), *натягивают* (1 раз), *рвут* (1 раз).

Неясно, трактовать ли каузацию движения (перемещение объекта) в качестве разновидности движения или нет. Если да, то ясно, почему в число глаголов движения входят глаголы каузации движения вроде *послать* (32 примера, которые можно в этом случае отнести к адекватным реакциям на исходный запрос, доведя тем самым общую степень адекватности ответов до 54%), но остаётся неясным, чем именно должен отличаться результат поиска по комбинации семантических признаков «движение» + «помещение

объекта» от запроса по одному из этих признаков («движение»). Если же нет, то неясно, почему глаголы типа *послать* трактуются как глаголы движения (*послать* 22 раза, *носить* 2 раза, *ниспослать* 1 раз, *отпустить* 1 раз, *отправить* 1 раз, *выпустить* 1 раз, *спустить* 1 раз, *привести* 1 раз, *принести* 1 раз, *привлекать* 1 раз). Единственным способом формулировки запроса о «движении, но не каузации движения» или «самостоятельном движении» является задание грамматического признака «непереходности» (обратим внимание, что, напр., глаголам *пересекать*, *покинуть*, *населить*, *заселить* признак «непереходности» не свойствен; поэтому использование в запросе дескриптора «непереходность» вместо «самостоятельного движения» будет означать, что конструкции с указанными глаголами в выдачу не попадут).

5. Примечательной особенностью ЛГП является то, что в нём не документирован поиск по отрицательным значениям признаков (грамматических и семантических). Для отчётливо бинарных оппозиций это не приводит ни к каким трудностям в формулировке запросов. Однако если оппозиция является более чем бинарной (как, например, грамматическая категория падежа или грамматический признак «часть речи»), то у неискушенного пользователя, читающего “help” к поиску по грамматическим признакам, может возникнуть впечатление, что в метаязыке запросов нет никаких формальных средств для задания таких запросов, как “–gen” или “–nom”, и потому некоторые типы запросов («с отрицанием») оказываются неосуществимыми. Напр., сформулировать запрос на поиск «сочетаний глагола *бояться* с последующим N, стоящим в форме “acc. sg.”, не являющейся одновременно при этом ни формой “nom.”, ни формой “gen.”, может лишь весьма искушенный пользователь. Ведь в “help”-е «по грамматическим признакам» ничего не сказано об отрицательном операторе (и, т. о., запись “acc &-gen &-nom” может построить лишь заведомо опытный пользователь).

6. Из комментариев к меню ВЫБРАТЬ неискушённый пользователь не может «по умолчанию» понять весьма важный момент, а именно, то, что результаты сделанного через это меню выбора разрешено исправлять “вручную”.