Национальный корпус русского языка и проблемы гуманитарного образования. Материалы международной научной конференции. Москва 19-20 апреля 2007. М. 2007. С. 77-79.

Использование Национального корпуса русского языка для решения задач моделирования речевой деятельности

Е.В.Ягунова
СПбГУ
iagounova_elena@mail.ru

В докладе планируется изложение результатов, полученных при решении ряда задач моделирования речевой деятельности с помощью Национального корпуса русского языка (НКРЯ).

1. **О выборе корпуса** современного русского языка

Строго говоря, этот пункт избыточен, т.к. на настоящее время НКРЯ является единственным корпусом современного русского языка, обладающим следующими характеристиками: (1) сплошная морфологическая аннотированность; (2) значительный объем корпуса (корпус со снятой омонимией насчитывает около 4 млн. с/у в 100 документах; нами использовался, главным образом, корпус со снятой омонимией); (3) доступность в качестве Интернет-ресурса для осуществления некоторых типов поиска, (4) возможность задавать тематические подкорпуса.

Последнее существенно для создания адекватных моделей процедур восприятия, включающих процедуры формирования «текущего словаря», которые отражают своего рода подстройку слушающего/читающего под лексико-семантические особенности воспринимаемого текста, что позволяет сузить рабочую область словаря. Соответственно облегчаются и становятся более эффективными процедуры идентификации (поиска в текущем словаре).

2. **О специфике рассматриваемых задач**

Необходимой частью моделирования речевой деятельности выступает область исследования, которая может быть названа ***лексикологией языка*** и ***лексикологией речевой деятельности*** (последний составной термин предложен В.Б. Касевичем в (Касевич 2005)). При моделировании ***восприятия*** речи существенно учитывать линейность (контактность) расположения единиц в тексте. Контактно могут располагаться как компоненты сложных словарных единиц, так и члены свободных словосочетаний. Именно здесь ценными могут оказаться данные корпуса. В частности, полезно обратиться к ***распределению частот встречаемости*** единиц и их типов, сочетаний (конструкций) и их типов. Высокая частота совместной встречаемости тех или иных слов дает основания для предположения о вхождении данного сочетания в словарь на правах отдельной единицы. Ср. представление об ***инвентарных*** и ***конструктивных*** единицах языка (Касевич 1988), где инвентарные единицы хранятся в

словаре, а конструктивные единицы порождаются в процессе речевой деятельности из инвентарных в результате применения правил.

3. В качестве **примера задачи** может служить исследование статуса разнообразных неоднословных целостностей (НЦ): фонетических слов (ФС), составных слов и сложных номинаций.

Напр., ***предложно-падежные конструкции с контактно расположенными компонентами*** (ср. *о_маме, к_маме)* обладают высокой фонетической и функциональной целостностью (предлоги здесь во многом «смыкаются» с приставками, особенно с омонимичными), что делает возможным считать конструкции целостными единицами словаря для слушающего. Высокая частотность предлогов в составе таких конструкций (см. примеры из табл.) может рассматриваться как существенный довод в пользу их предположительной «инвентарности».

Таблица. Некоторые предлоги, омонимичные частотным приставкам (НКРЯ)

	Всего	С контактными	
		сущ./мест.-сущ.	сущ.
за	ок. 150 тыс.	Ок. 100 тыс.	ок. 80 тыс.
на	ок. 500 тыс.	Ок. 300 тыс.	ок. 250 тыс.
с	ок. 350 тыс.	Ок. 200 тыс.	ок. 200 тыс.
у	ок. 200 тыс.	Ок. 150 тыс.	ок. 50 тыс.
в	ок. 500 тыс.	Ок. 350 тыс.	ок. 350 тыс.
по	ок. 300 тыс.	Ок. 200 тыс.	ок. 200 тыс.

Наряду с имеющимися возможностями решения поставленной задачи отметим ряд параметров, препятствующих полноте ее рассмотрения: (1) неточный подсчет частот (для сравнительно большой частоты встречаемости) средствами поиска Яндекса; (2) отсутствие учета знаков препинания искажает статистику, включая в результаты поиска и конструкции со знаками препинания внутри; (3) отсутствие возможности создания частотного словаря, в т.ч. по результатам поиска пользователь имеет возможность получить только набор контекстов, но не частотный словник.

Последнее необходимо, напр., при поиске наиболее частотных конструкций (словосочетаний), при сопоставлении омонимичных пар наподобие *от того – оттого; с начала – сначала; в конец – вконец* и т.д. На настоящий момент рассмотрение такого типа вопросов осуществлялось нами в два этапа: (1) на основании распределения частот встречаемости по «Генеральному корпусу» («Г.К.») (созданному С.А. Крыловым на основе Уппсальского корпуса) составляются списки потенциальных кандидатур в инвентарные единицы; (2) эти списки верифицируются с помощью поиска по НКРЯ.

Другой пример исследования неоднозначности касается омонимии «свободное сочетание – составное слово» (напр., *друг друга* и *может быть)*. Для *друг друга* очевидно статистическое предпочтение в пользу составного

слова: в НКРЯ не было найдено ни одного контекста со свободным словосочетанием. Иначе обстоит дело с более ровным соотношением частот в случае омонимии *мо+жет_быть* : *мо+жет бы+ть* (ударение «+»), т.к. выбор между вариантами не может осуществляться на основании статистического критерия. Вероятным критерием выбора является синтаксический и, соответственно, пунктуационный. Для рассмотрения данного типа омонимии мы обращались к частоте встречаемости по НКРЯ. Однако это сочетание достаточно частотно (516 документов, ок. 2 тыс. контекстов), поиск с учетом знаков препинания невозможен, следовательно, решение вопроса возможно, но требует «ручного» просмотра всех контекстов.

Естественно, «ручной» просмотр всех контекстов из НКРЯ – единственный способ рассмотрения вопросов, связанных с омонимией в рамках потенциально сложных номинаций: напр., *большой театр*, где параметр «прописная/строчная» напрямую не работает, а наряду со значением имени собственного может идти речь (1) о физическом размере здания театра, его труппе и т.п.; (2) о соответствующем виде искусства (ср. *большое искусство*) или (3) компонент *театр* входит в более сложное наименование (наподобие *театра кукол, театра военных действий* и т.п.).