Н. В. Козловская (Санкт-Петербург)
mnegolosbyl@gmail.com
А. Ю. Козловский (Санкт-Петербург)
alexander.kozlovsky@gmail.com

ЧЕТЫРЕ ИМЕНИ ИМПЕРАТОРА: КОНКОРДАНС И «ОБЛАКО СЛОВ» КАК МЕТОДЫ ОТБОРА ИМЕН СОБСТВЕННЫХ В ДИФФЕРЕНЦИАЛЬНЫЙ АВТОРСКИЙ СЛОВАРЬ

В докладе ставится проблема отбора имен собственных из текста художественного произведения с целью лексикографического представления в дифференциальном объяснительном словаре военной лексики романа «Война и мир». По предварительным данным, в произведении использовано более 700 имен собственных.

Задача начального этапа – составление рабочего словника имен собственных с учетом содержания и параметров разрабатываемого словаря. Представляется возможным использование нескольких способов отбора: традиционный, корпусный (и их сочетание), а также «облако слов» как аналитический инструмент.

Сплошная выборка как традиционный метод составления словника представляется излишне трудоемким. Поскольку текст романа является объектом разноаспектного филологического анализа в течение долгого времени, списки имен собственных введены в научный обиход: так, на портале tolstoy-lit.ru представлен «указатель собственных имен», составленный З.Н. Ивановой. В этот список входят топонимы разных типов, обозначения исторических событий, названия книг, различных произведений искусства и пр.

В указатель не включен параметр частоты, однако количество номеров страниц к каждому имени собственному косвенно указывает на количество употреблений слова в романе (без конкретных показателей). Отметим также, что из четырех имен императора Франции в указатель включены два: *Наполеон* и *Бонапарт*.

Анализ опубликованного указателя в сочетании с методом конкорданса на базе пользовательского подкорпуса в НКРЯ или Sketch Engine позволяет сформировать круг имен собственных, подлежащих анализу и описанию.

Например, согласно данным НКРЯ, имя собственное *Штраух* встречается в окончательной редакции романа только один раз в контексте, который трудно назвать репрезентативным и интересным: *В то время как князь Андрей сошелся с Несвицким и Жерковым, с другой стороны коридора навстречу им шли Штраух, австрийский генерал, состоявший при штабе Кутузова для наблюдения за продовольствием русской армии, и член гофкригсрата, приехавшие накануне*. [Л. Н. Толстой. Война и мир / Том первый (1869)].

Второй путь извлечения имен собственных для включения в словник – обращение к аналитическому инструменту Wordlist (список частот) в системе Sketch Engine без заданных специальных частотных характеристик. Наиболее частотные названия и имена подлежат проверке на соотнесенность с военной

концептосферой текста и включению или невключению в словник. Первым по критерию частоты в многостраничном списке является имя персонажа *Андрей* (1402). Следующие позиции в списке (за исключением имен персонажей) занимают слова: *Наполеон* 712, *Кутузов* 609, *Александр* 229, *Багратион* 95, *Бонапарте* 78, *Аустерлицкое* (сражение) 26, *Аустерлиц* 25, *Бонапарт* 9 (?).

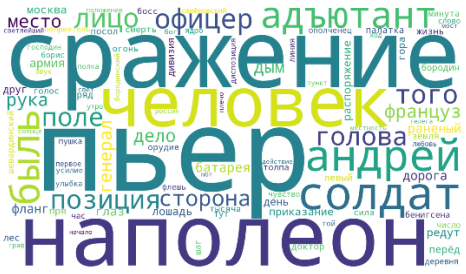
Автоматическим образом полученные данные требуют комментариев и лингвистической интерпретации. Заметим также, что статистические данные Sketch Engine и НКРЯ довольно сильно разнятся, что также требует размышлений и уточнений.

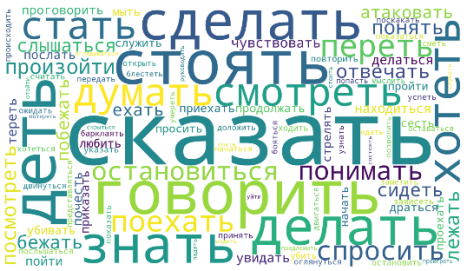
Помимо традиционных корпусных технологий, статистическую оценку текста и первоначальную выборку слов можно осуществлять с помощью инструментов компьютерных библиотек. В поисках инструментов по автоматическому выявлению имен собственных в тексте авторы доклада провели небольшой эксперимент по созданию «облака слов» к конкретным текстовым фрагментам.

Для эксперимента были выбраны три фрагмента: полные описания Шенграбенского и Бородинского сражений, а также блок философских рассуждений Толстого о философии истории в конце романа.

Вначале текст был обработан с помощью библиотеки pymorphy2, позволяющей определить начальную форму каждого слова в тексте, его частеречную принадлежность и другие грамматические характеристики. Это послужило основой для дальнейшего распределения слов по заданным категориям (существительные, прилагательные, глаголы) и создания частотного словаря слов. Следующим шагом стала фильтрация слов по длине и исключение широко употребляемых слов и предлогов, которые могут исказить общую картину анализа. На основе полученного частотного словаря формируется облако слов (количество выбранных единиц – 100) с помощью библиотеки wordcloud, которая позволяет настроить визуальное представление облака слов, включая цвет фона, максимальное количество слов и размер изображения. Визуализация происходит средствами библиотеки matplotlib, широко применяемой для создания графиков и изображений в научных исследованиях.

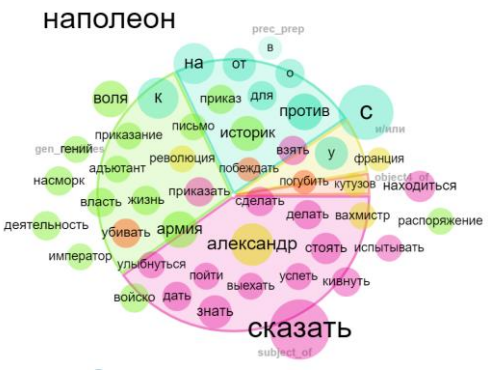
Поскольку предметом настоящего доклада являются только имена собственные, остановимся на анализе «облаков существительных». В основе эксперимента лежит предположение о том, что в «облако существительных» батального эпизода войдут наиболее значимые имена собственные, которые подлежат включению в словник. Наиболее значимы немногочисленные имена собственные: *Наполеон* (размер шрифта визуализирует и подчеркивает значимость этой единицы в лексической структуре эпизода), *Москва*, *Шевардинский* (редут), *Бенигсен*, *Бородино*, *Семеновский*. Схема неидеальна: очевидно, что программа не опознала фамилию *Кутузов* как существительное и «вывела» ее за пределы облака.





Сделаем частное наблюдение, которое представляется важным и интересным. В «облаке глаголов» Бородинского сражения число глаголов, обозначающих действия воюющих, дерущихся, невелико: *убивать, стрелять, драться, атаковать* и некоторые другие. Отметим также наличие «шума» в полученных данных: прилагательное *бенигсный*, глагол *барклаять*, что говорит о несовершенстве аналитических инструментов системы. Это определяет направление приложения дальнейших усилий программиста (коррекция заданных характеристик) и лингвиста (увеличение количества слов в «облаке»).

На этапе составления словарной статьи целесообразно использовать «словесные эскизы» системы Sketch Engine, отражающие ключевые слова, группирующиеся вокруг имен собственных. Многие глагольные сочетания отражают концептосферу войны, но самым частотным является глагол речи: *говорить, сказать*.



Анализ текстового указателя имен, данных корпусов и экспериментальных «облаков слов», сгенерированных традиционными программными инструментами, говорит о том, что доступных, известных методов автоматического выявления из текста имен собственных – по заданным характеристикам – не существует. В связи с этим авторами статьи ставится задача разработки программы, облегчающей данную задачу. Ни один из предложенных методов не позволил обнаружить все четыре имени, включая французский вариант написания. В настоящее время найти «четыре имени французского императора» может только человек, внимательно читающий и / или хорошо знающий текст.