ПРОБЛЕМЫ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ТЕКСТОВ

Г. Б. Байраммырадов (*Гродно, ГрГУ*)

ИСПОЛЬЗОВАНИЕ ИНСТРУМЕНТА N-ГРАММ НАЦИОНАЛЬНОГО КОРПУСА РУССКОГО ЯЗЫКА ДЛЯ ПОИСКА УСТОЙЧИВЫХ СОЧЕТАНИЙ С ЭТНОНИМАМИ

В статье рассматриваются возможности поиска устойчивых сочетаний с этнонимами *немец / немка* и *англичанин / англичанка* с помощью инструмента 5-граммы в Национальном корпусе русского языка.

Различного рода составные объекты (multi-word expressions – MWEs) являются предметом исследования во многих направлениях современной лингвистики. Наиболее исчерпывающий перечень типов составных объектов, не имеющих отношения к терминологии, был дан Л. В. Рычковой в [1]. Ею же была определена важность использования корпусных технологий для идентификации такого рода объектов как на основе использования методов лингвостатистики, так и фреймовых или модельных структур [1, с. 191]. Еще одним инструментом, позволяющим осуществлять поиск устойчивых сочетаний – потенциальных составных объектов, – является поиск n-грамм в корпусах.

N-граммы – это инструмент, с помощью которого осуществляется поиск комбинаций, состоящих из нескольких слов, «из подкорпуса с неснятой омонимией основного корпуса» Национального корпуса русского языка (НКРЯ). Такой инструмент позволяет искать и точные формы, и леммы слов «с учетом грамматических признаков и пунктуации или без». Результат поиска позволяет получить информацию о частоте комбинаций с искомым словом и о количестве содержащих их документов [2].

В НКРЯ возможен поиск данных по биграммам, триграммам, 4-граммам и 5-граммам. Подробное описание данной процедуры для каждой n-граммы отражено в инструкции (рис. 1).

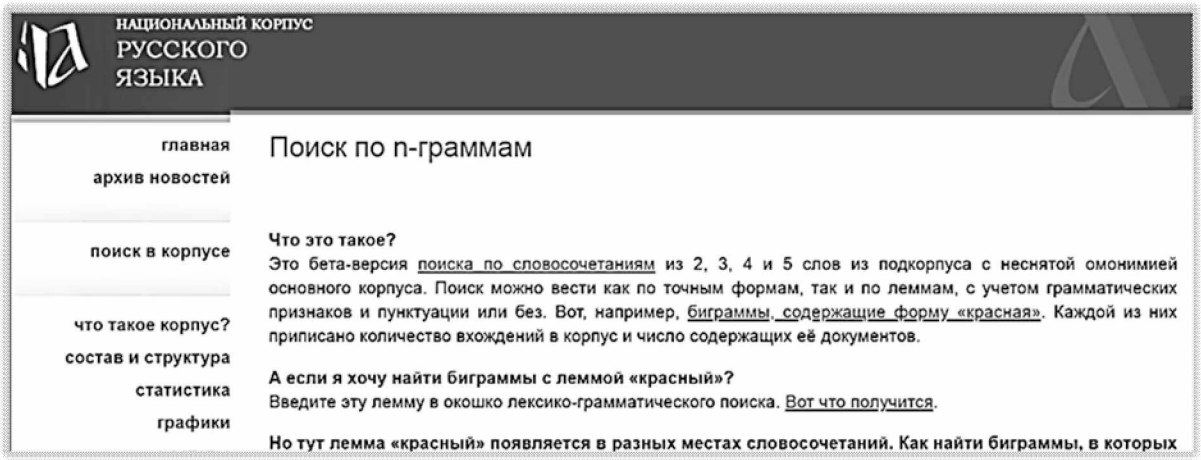
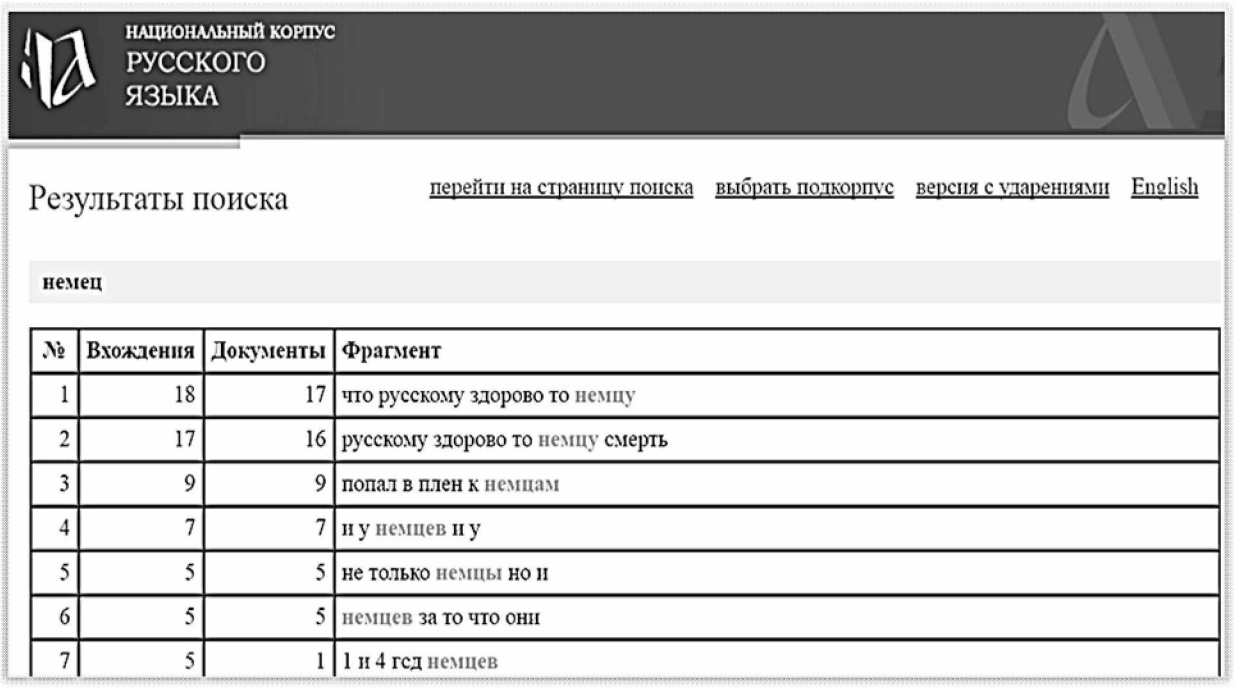


Рис. 1. Инструкция по поиску n-грамм

Следует отметить, что при осуществлении поиска с применением инструмента n-грамм нельзя использовать возможности семантической разметки, поэтому одновременное осуществление поиска этнонимов с помощью семантической разметки и комбинаций с ними с применением инструмента n-грамм не представляется возможным.

Таким образом, использование n-грамм предполагает изначальное определение конкретных этнонимов, поиск комбинаций с которыми и позволит выявить устойчивые сочетания. Для иллюстрации возможностей инструмента n-грамм были выбраны этнонимы *немец / немка* и *англичанин / англичанка*, а также 5-граммы как наиболее информативные с точки зрения объема получаемых контекстов. Применение опции лексико-грамматического поиска позволяет получить комбинации со всеми грамматическими формами искомых этнонимов.

Поиск 5-грамм с этнонимом *немец* позволил получить 100 комбинаций, из них 80 имеют частотность 3 и выше. В полученной нами выдаче первым, самым частотным, оказался фразеологизм *что русскому здорово, то немцу смерть*. Под фразеологизмом, вслед за А. Н. Тихоновым, мы понимаем воспроизводимую в речи единицу, имеющую «целостное значение, постоянный компонентный состав и грамматическую структуру» [3, с. 116–117]. Данный фразеологизм состоит из более чем 5 слов; это доказывает, что с помощью инструмента 5-грамм можно найти устойчивые сочетания с большим количеством слов (см., например, комбинации 1 и 2 на рис. 2).


НАЦИОНАЛЬНЫЙ КОРПУС РУССКОГО ЯЗЫКА

Результаты поиска

перейти на страницу поиска | выбрать подкорпус | версия с ударениями | English

немец

№	Вхождения	Документы	Фрагмент
1	18	17	что русскому здорово то немцу
2	17	16	русскому здорово то немцу смерть
3	9	9	попал в плен к немцам
4	7	7	и у немцев и у
5	5	5	не только немцы но и
6	5	5	немцев за то что они
7	5	1	1 и 4 гсд немцев

Рис. 2. Фрагмент выдачи с фразеологизмом, состоящим из более чем 5 слов

Среди найденных комбинаций с этнонимом *немец* самыми частотными оказались следующие словосочетания: *попасть в плен к немцам* (9 вхождений в 9 документах – см. рис. 2) и *опорный пункт обороны немцев* (см. комбинации под номерами 8, 11, 14, 15, 16, 17, 18, 19 на рис. 3).

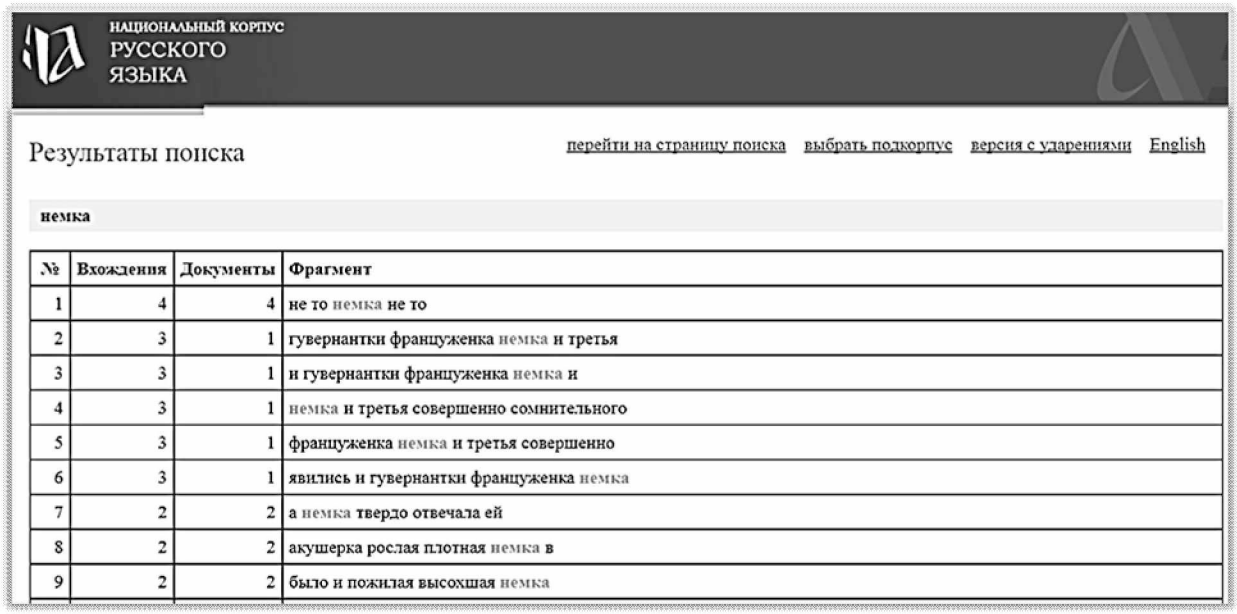
8	4	4	важным опорным пунктом обороны немцев
9	4	4	и французы и немцы и
10	4	4	мама и убитый немцами вечер
11	4	4	мощным опорным пунктом обороны немцев
12	4	4	народ рабами немцев я не
13	4	4	не то немец не то
14	4	4	обороны немцев на реке морава
15	4	4	опорным пунктом обороны немцев в
16	4	4	опорным пунктом обороны немцев на
17	4	4	пунктом обороны немцев в саксонии
18	4	4	пунктом обороны немцев на реке
19	4	4	сильными опорными пунктами обороны немцев

Рис. 3. 5-граммы, содержащие словосочетание *опорный пункт обороны немцев,* в различных грамматических формах

Ожидаемо частотными оказались фразеосхемы *не только X, но и Y* , где вместо X употребляется этноним *немцы*, а вместо Y – другой этноним. Среди других частотных комбинаций – *немцев за то, что они сделали*; *и французы и немцы / и немцы и французы*; *российские немцы*; *сделать русский народ рабами немцев*; *не считать за немцев*; *в тыл к немцам* и *в тылу у немцев*; *война с немцами*; *не видеть ни одного немца* и (несколько неожиданные на общем фоне) *болеть / рад за немцев*.

Применение инструмента 5-грамм позволило также найти перефразированную детскую считалку *(вышел немец из тумана, вынул ножик из кармана)*, пословицу *(немец обезьяну выдумал*), название стихотворения В. В. Маяковского *(«Мама и убитый немцами вечер»)*. При использовании инструмента n-грамм, особенно 5-грамм, вместо знаменательных слов в комбинациях могут зачастую встречаться, например, служебные слова или числа: *и у немцев и у*; *1 и 4 гсд немцев*; *не то немец не то*, *не у немцев не у* и др. Отметим, что разработчики НКРЯ заранее предусмотрели возможность исключения знаков препинания при выдаче компонентов n-грамм.

Поиск 5-грамм с этнонимом *немка* позволил найти только одно устойчивое сочетание, поскольку из полученных в его результате 5-грамм только одна комбинация, использующая фразеосхему *не то X не то Y, – не то немка, не то* – имеет высокую встречаемость, а остальные комбинации можно считать шумовыми, либо они имеют низкую встречаемость (рис. 4).



НАЦИОНАЛЬНЫЙ КОРПУС РУССКОГО ЯЗЫКА

Результаты поиска

перейти на страницу поиска | выбрать подкорпус | версия с ударениями | English

немка

№	Вхождения	Документы	Фрагмент
1	4	4	не то немка не то
2	3	1	гувернантки француженка немка и третья
3	3	1	и гувернантки француженка немка и
4	3	1	немка и третья совершенно сомнительного
5	3	1	француженка немка и третья совершенно
6	3	1	явились и гувернантки француженка немка
7	2	2	а немка твердо отвечала ей
8	2	2	акушерка рослая плотная немка в
9	2	2	было и пожилая высохшая немка

Рис. 4. Результат поиска 5-грамм с этнонимом *немка*

Поиск с этнонимом *англичанин* показал, что самым частотным в корпусе оказался контекст *Англия и англичане – с американской точки зрения*, затем, по убыванию частотности, – *с тех пор как англичанин; ни англичанин, ни француз, ни немец; и французы и англичане и…* В нескольких документах встретились фрагменты из произведения Ф. М. Достоевского «Подросток»: *останется наиболее французом, равно англичанин и немец*; *никогда, никогда, никогда англичанин не будет рабом* (рис. 5).

№	Вхождения	Документы	Фрагмент
1	6	1	англичане с американской точки зрения
2	6	1	англия и англичане с американской
3	6	1	и англичане с американской точки
4	5	5	с тех пор как англичане
5	3	3	англичанин ни француз ни немец
6	3	3	и французы и англичане и
7	3	3	наиболее французом равно англичанин и
8	3	3	ни англичанин ни француз ни
9	3	3	останется наиболее французом равно англичанин
10	3	3	французом равно англичанин и немец
11	3	2	никогда англичанин не будет рабом
12	3	2	никогда никогда англичанин не будет
13	3	2	никогда никогда никогда англичанин не
14	3	1	вовсе не означает что англичане
15	2	2	xi после изгнания англичан при

Рис. 5. Выдача 5-грамм с этнонимом *англичанин*

Самое большое количество комбинаций из всех полученных выдач приходится на 5-граммы с этнонимом *англичанка.* Тем не менее все 189 полученных комбинаций нельзя отнести к устойчивым, так как их частотность ниже 3 (рис. 6).

№	Вхождения	Документы	Фрагмент
1	2	2	англичанка гувернантка маленькое худенькое существо
2	2	2	англичанками учительницами английского языка в
3	2	2	англичанку с которой была списана
4	2	2	англичанку что он в молодости
5	2	2	благодарность ловкая гимнастка-англичанка прыгала к
6	2	2	в благодарность ловкая гимнастка-англичанка прыгала
7	2	2	в таинственный дар упомянутых англичанок
8	2	2	гимнастка-англичанка прыгала к нему на
9	2	2	двумя англичанками учительницами английского языка
10	2	2	жила англичанка гувернантка маленькое худенькое
11	2	2	красивого типа женщин как англичанки

Рис. 6. 5-граммы с этнонимом *англичанка*

Проведенный анализ показывает, что инструмент n-грамм позволяет получать сочетания, регулярно встречающиеся вместе в письменной речи [4]. Наличие среди полученных выдач фразеологизмов, пословиц и частотных словосочетаний дает возможность выявить маркированность одних этнонимов в русской лингвокультуре и нейтральность других. Так, явной негативной маркированностью отличается этноним *немцы* и несколько меньшей – *немец*. Это особенно ярко видно на фоне этнонимов *англичане/англичанин*, являющихся нейтральными. Нейтральными являются и этнонимы женского рода.

Таким образом, применение инструмента n-грамм при введении в процесс обучения этнонимов отвечает целям иноязычного образования, а использование возможностей этого инструмента при обучении русскому языку иностранцев позволяет сочетать освоение ими семантической и синтаксической валентностей с изучением сложных предлогов, фразеосхем, фразеологизмов, паремий, а также мотивировать к чтению русской художественной литературы и познанию истории для того, чтобы найти объяснение маркированности определенных этнонимов в русской лингвокультуре.

ЛИТЕРАТУРА

1. *Рычкова, Л. В.* Праблема састаўных аб'ектаў у корпусах славянскіх моў і лінгвістычных базах дадзеных / Л. В. Рычкова // Мовазнаўства. Літаратура. Культуралогія. Фальклярыстыка: XIII Міжнар. З'езд славістаў: дакл. бел. дэлегацыі. Любляна, 2003. – Мінск : Беларус. навука, 2003. – С. 184–195.

2. Национальный корпус русского языка [Электронный ресурс]. – Режим доступа : www.ruscorpora.ru. – Дата доступа : 22.04.2018.

3. *Тихонов, А. Н.* Современный русский язык. Теоретический курс. Лексикология / А. Н. Тихонов. – М. : Рус. яз., 1987. – Ч. 2 – С. 115–136.

4. *Rumer, U.* The inseparability of lexis and grammar: Corpus linguistic perspectives / U. Rumer // Annual Review of Cognitive Linguistics. – 2009. – Vol. 7. – P. 141–163.

The article considers possibilities of search of the stable word combinations containing ethnonyms using the 5-grams tool provided by the National Corpus of the Russian language.