categories in the choice of the description's subject, i.e. important socio-psychological qualities for a given age group; and interpersonal connections, i.e. unidirectional and bi-directional (reciprocal) connections which help to identify the leader and outsider positions in the class.

Key words: linguistic portrait, semantic field, categorization, interpersonal relations.

УДК 81'35

ЗНАЧИМАЯ ЛЕКСИКА ОФИЦИАЛЬНОГО ДОКУМЕНТА НА ФОНЕ ДВУХ РЕФЕРЕНТНЫХ КОРПУСОВ[1]

Блинова О. В.

Санкт-Петербургский государственный университет, 199034, Россия, г. Санкт-Петербург, Университетская наб., 7/9, o.blinova@spbu.ru

Статья направлена на сравнение списков значимой лексики (resp. списков ключевых слов), полученных с использованием двух разных по содержанию референтных корпусов одинакового объёма (по 1 млн. слов). Целевой корпус – это лемматизированнная субколлекция текстов русских локальных документов типа «Информированное добровольное согласие на медицинское вмешательство» (38,5 тыс. слов). Референтными корпусами являются офлайновый версии НКРЯ и ГИКРЯ со снятой грамматической омонимией. Исследование показало, что «верхушки» двух списков «положительных» ключевых слов очень схожи, а сами ключевые слова отражают прежде «о-чёмность» («aboutness») текстов.

Ключевые слова: официальные документы, корпус русских локальных документов CorRIDA, значимая лексика, ключевые слова, референтный корпус.

1. В статье апробируется одна из методик выявления значимой лексики, которая в дальнейшем может применяться при анализе текстовых коллекций документов, вошедших в корпус CorRIDA (Corpus of Russian Internal Documents and Acts, Корпус русских локальных документов и актов).

Корпус CorRIDA формируется в рамках исследования, посвящённого функционированию официальных документов в социальных доменах здравоохранения, культуры и образования. В корпус включены так называемые «локальные документы» (Internal Documents) конкретных государственных учреждений, находящиеся в открытом доступе на сайтах поликлиник, школ, университетов, библиотек и др. Это тексты, с которыми мы как «простые» носители языка часто сталкиваемся в повседневной жизни. Конечной

[1] Исследование выполнено в рамках НИР по анализу соблюдения норм современного русского литературного языка при его использовании в качестве государственного в деятельности организаций культуры, здравоохранения и образования, включённой в план мероприятий НИИ Проблем государственного языка СПбГУ во исполнение комплекса мер, направленных на совершенствование государственной политики в области развития, защиты и поддержки русского языка на 2016-2020 гг. Совета по русскому языку при Правительстве РФ.

целью исследования является оценка языка документов с точки зрения доступности для восприятия и понимания.

В настоящей статье анализируется одна коллекция из полуторамиллионного корпуса CorRIDA, это коллекция «информированных согласий», которые функционируют в домене здравоохранения («Информированное добровольное согласие на медицинское вмешательство» и др.). В коллекцию входит 100 текстов информированных согласий объёмом 50,6 тыс. токенов, 38,5 тыс. слов[2].

2. Списки значимой лексики (или, иначе, списки ключевых слов, keywords) выявляются путём сравнения частот слов в двух корпусах. Первый корпус называется «целевым», или «объектным» (target corpus, далее – ЦК); его лексическая специфика выявляется «на фоне» второго, «референтного», корпуса (reference corpus, далее – РК). В результате получается список собственно ключевых слов, positive keywords, которые встречаются в ЦК значимо чаще, чем можно было бы ожидать, если судить по РК, а также список «негативных» ключевых слов, negative keywords, чьи частоты значимо низки в сравнении с РК. Метод выделения ключевых слов позволяет выявить лексическую специфику ЦК по сравнению с РК.

РК обычно более «общий» и «сбалансированный», а целевой – более «специализированный» [1]. Кроме того, считается, что РК должен быть (существенно) больше целевого. Например, в работе [2] указывается, что если РК больше ЦК менее, чем в пять раз, то использовать его рискованно. Между тем Э. МакЭнери и Р. Сяо [3], сравнивая два РК: 100-миллионный BNC и 1-миллионный Freiburg-LOB Corpus, получили практически идентичные списки ключевых слов, после чего заключили, что размер референтного корпуса не так уж и значим. М. Скотт также считает, что состав РК важнее размера (и при использовании разных жанрово-специфичных РК получаются существенно различающиеся списки ключевых слов) [4].

3. Рассмотрение ключевых слов широко используется для целей дискурсивного анализа, контент-анализа, изучения стилей и др. (обзор см., например, в работе [5]). Анализ ключевых слов, в частности, позволяет идентифицировать существенные особенности текстов, связанные с их жанровой спецификой [3, с. 68].

Существуют инструменты автоматического извлечения ключевых слов: WordSmith Tools, AntConc, WMatrix [5, с. 95]. В настоящем исследовании для этой цели используется программный пакет AntConc [6]. При подсчёте коэффициента keyness value в качестве метрики используется критерий отношения правдоподобия.

[2] Из списка слов, подлежащего анализу, исключены цифровые и некоторые алфавитно-цифровые комплексы, пунктуационные символы, в том числе числовые и буквенные обозначения пунктов документа.

Таблица 1. Списки «положительных» ключевых слов, полученных с помощью двух РК

НКРЯ как РК				ГИКРЯ как РК			
№	**Абс. частота**	**Keyness**	**Ключевое слово**	**№**	**Абс. частота**	**Keyness**	**Ключевое слово**
1	855	5092.517	медицинский	**1**	855	5200.907	медицинский
2	442	2626.464	лечение	**2**	442	2489.885	лечение
3	360	2146.122	согласие	**3**	360	2204.359	согласие
4	402	2064.088	врач	**4**	402	1959.546	врач
5	258	1658.182	информировать	**5**	258	1659.369	информировать
6	242	1491.317	подпись	**6**	240	1477.739	вмешательство
7	240	1438.399	вмешательство	**7**	242	1382.529	подпись
8	241	1265.169	здоровье	**8**	204	1379.029	они
9	240	1264.638	проведение	**9**	188	1270.870	давать
10	287	1252.489	г	**10**	196	1258.003	ф

Задачей является извлечение списков ключевых слов с использованием двух разных РК и сравнение таких списков. В качестве РК взяты офлайновая версия подкорпуса Национального корпуса русского языка (НКРЯ, ruscorpora.ru) со снятой грамматической омонимией и офлайновая версия подкорпуса Генерального интернет-корпуса русского языка (ГИКРЯ, webcorpora.ru) со снятой грамматической омонимией[3]. Решено использовать два РК равного объёма (по 1 млн. слов). Подкорпус НКРЯ включает тексты электронной коммуникации, художественную и учебно-научную прозу, газетную публицистику, материалы устной речи (1 млн 61 тыс. слов). Подкорпус ГИКРЯ – это тексты интернет-коммуникации «ВКонтакте» и «ЖЖ», взято 1 млн. 92 тыс. слов[4]. ЦК, состоящий из текстов официальных документов («информированных согласий») существенно (в 25 раз) меньше референтного (38,5 тыс. слов).

Перед формированием списков ключевых слов словник ЦК был лемматизирован с помощью MyStem [7]. Кроме того, для унификации списков лемм были выполнены некоторые дополнительные преобразования, в частности, в составе списка лемм ГИКРЯ все объединённые нижним подчеркиванием неоднословные лексические единицы типа *в_течение* были разбиты на части.

[3] Пользуясь случаем, хочу ещё раз поблагодарить коллективы НКРЯ и ГИКРЯ за предоставление доступа к офлайн-корпусам.

[4] В состав РК вошли в основном тексты ЖЖ («Живого журнала»).

Таблица 2. Списки «отрицательных» ключевых слов, полученных с помощью двух РК

НКРЯ как РК				ГИКРЯ как РК			
№	**Абс. частота**	**Keyness**	**Ключевое слово**	**№**	**абс. частота**	**Keyness**	**Ключевое слово**
1	2	0.000	неизвестный	**1**	1	0.000	методика
2	3	0.000	январь	**2**	1	0.000	нормативный
3	6	0.000	участие	**3**	1	0.000	полоса
4	1	0.001	механический	**4**	1	0.000	частый
5	1	0.001	речевой	**5**	2	0.000	кв
6	1	0.001	слабость	**6**	2	0.001	праздничный
7	1	0.001	транспортный	**7**	6	0.002	свободный
8	3	0.001	корень	**8**	2	0.003	влиять
9	10	0.001	правый	**9**	1	0.003	запрос
10	28	0.001	проблема	**10**	6	0.003	способный

В работе [3] учёные, получив лексические списки с помощью двух РК, сравнивали 10 «позитивных» ключевых слов («top ten positive keywords») и 10 «негативных» ключевых слов («top ten negative keywords»). Основания подобного сравнения для наших данных представлены в табл. 1 и 2 ниже (списки отсортированы по убыванию коэффициента «keyness» в табл. 1 и по возрастанию этого коэффициента в табл. 2).

Первые пять позиций совпадают и по составу ключевых слов, и по месту этих слов в списке (это *медицинский, лечение, согласие, врач, информировать*). Вторые пять позиций содержат два совпадения с разным порядком следования (*подпись* и *вмешательство*). Стоит прокомментировать наличие на 10-й позиции начальной буквы слова *ф* (из акронима *Ф. И. О.*) и начальной буквы слова *г.* (это сокращение слова *год*). Таким образом, в текстах документов по сравнению с референтными корпусами значимо часто встречаются сокращения.

Можно сделать некоторые предварительные выводы: во-первых, списки «положительных» ключевых слов, полученные с помощью разных РК, существенно схожи (совпадают 7 позиций из 10, первые 5 позиций совершенно идентичны); во-вторых, списки «негативных» ключевых слов не содержат совпадений.

Отмечено, что ключевые слова – это чаще всего: 1) имена собственные; 2) слова, которые и люди (а не только компьютерные программы) посчитали бы ключевыми, такие единицы являются хорошими показателями

тематики текста; 3. слова, которые являются индикаторами стилевых характеристик текстов, а не их тематики [8]. В полученных списках «положительных» ключевых слов преобладают единицы второй категории, говорящие, «о чём текст».

Библиографический список

1. *Gries S.T.* Quantitative designs and statistical techniques // The Cambridge handbook of English corpus linguistics / ed. by D. Biber and R. Reppen. Cambridge University Press. 2015. P. 50-71.
2. *Berber-Sardinha T.* Comparing corpora with WordSmith Tools: How large must the reference corpus be? // Proceedings of the Workshop on Comparing Corpora. 2000. Vol.9. P. 7-13.
3. *Xiao R., McEnery T.* Two approaches to genre analysis: Three genres in modern American English // Journal of English Linguistics. 2005. Vol. 33. P. 62–82.
4. *Scott M.* In search of a bad reference corpus // What's in a word-list? Investigating word frequency and keyword extraction / ed. by D. Archer. Farnham, Surrey: Ashgate. 2009. P. 79-91.
5. *Culpeper J., Demmen J.* Keywords // The Cambridge handbook of English corpus linguistics / ed. by D. Biber and R. Reppen. Cambridge University Press. 2015. P. 90-105.
6. *Anthony L.* AntConc (Version 3.5.6). Tokyo, Japan: Waseda University. URL: http://www.laurenceanthony.net/software (дата обращения: 01.04.2018).
7. Морфологический анализатор Mystem 3.*1* URL: https://tech.yandex.ru/mystem/ (дата обращения: 01.04.2018).
8. Definition of key-ness. Wordsmith Tools online manual. URL: http://www.lexically.net/wordsmith/index.html (дата обращения: 01.04.2018).

OFFICIAL DOCUMENTS' KEYWORDS LISTS MADE ON THE BASE OF TWO REFERENCE CORPORA: A COMPARISON

Blinova O. V.

St. Petersburg State University, 7/9, Universitetskaya emb., St. Petersburg, 199034, Russia, o.blinova@spbu.ru

The article is aimed at comparing the lists of keywords, obtained using two different reference corpora of the same size (1 million words). The target corpus is a lemmatized collection of texts of internal documents such as «The informed voluntary consent to medical intervention» (38,5 thousand words). The reference corpora are the off-line disambiguated subcorpus of the RNC (ruscorpora.ru) and the off-line disambiguated subcorpus of the GICR (webcorpora.ru). The study showed that the «tops» of two lists of positive keywords are very similar, and the keywords themselves indicate primarily «aboutness» of the texts.

Key words: official texts, Corpus of Russian Internal Documents and Acts «CorRIDA», keywords, reference corpus.