© А.Ю. ШЕВЦОВА
shevcova05@mail.ru

УДК 81'33

МЕТОДЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ В РАЗРАБОТКЕ ТИПОЛОГИЧЕСКИХ ПАРАМЕТРОВ НОРМАТИВНОГО ДВУЯЗЫЧНОГО СЛОВАРЯ-ТЕЗАУРУСА ТЕРМИНОВ АВИАЦИОННОЙ ЭКОЛОГИИ

APPLYING METHODS OF MATHEMATICAL STATISTICS TO DEFINE TYPOLOGICAL CHARACTERISTICS OF THE PRESCRIPTIVE ECOLOGICAL THESAURUS OF CIVIL AVIATION

АННОТАЦИЯ. Автор предпринимает попытку представить и описать методику отбора терминов для формирования и обоснования терминологического массива нормативного двуязычного словаря-тезауруса терминов авиационной экологии. В разработке принципов отбора терминологического массива моделируемого словаря были использованы методы математической статистики, а именно критерий Пирсона хи-квадрат. Используя данный критерий, последовательно сравнивали пять выборок терминов, сформированных путем отбора из нормативных источников исследуемой области знания. Выборки сравнивали дважды для подтверждения количественного и качественного сходства между ними, а следовательно, сравнивали два вида показателей: количественные, выраженные в частотном распределении терминов в корпусе научных текстов, и качественные, выраженные в соотношении научных и технических терминов в выборках. Проведенный статистический анализ выборок показал, что выборочная совокупность терминов является на 70% однородной по количественным показателям и на 60% изоморфной по качественным показателям, а значит, может считаться принадлежащей одной генеральной совокупности, и быть представленной в одном терминологическом словаре. Отобранный терминологический массив находится в рамках нормативной парадигмы исследуемой предметной области и определяет нормативную функцию в качестве доминирующей функции моделируемого словаря. При этом словарная функция нормативности закладывается как основной композиционный типологический параметр. Таким образом, метод математической статистики обеспечил отбор терминов в нормативный двуязычный словарь-тезаурус и определил типологические параметры этого словаря.

SUMMARY. The article describes the methodology of terms' selection for forming and validating terminological corpus of the prescriptive ecological thesaurus of civil aviation. The methods of mathematical statistics, Pearson criterion chi-square in particular, have been used in order to develop the principles of terminological corpus selection. With the help of the criterion we consistently compared five samplings of the terms taken from regulatory documents of the domain under consideration. We compared the samplings

twice to confirm their quantitative and qualitative isomorphism. In other words, two types of characteristics were compared: quantitative based on terms' frequency distribution in the corpus of scientific texts, and qualitative based on correlation of scientific and technical terms in the samplings. The statistical analysis of the samplings proved that the terms' selection is quantitatively homogeneous by 70% and qualitatively isomorphic by 60%. So, the selection belongs to one entire assembly and may be included in one terminological dictionary. The selected terminological corpus lies within the regulatory paradigm of the considered domain and determines a prescriptive function as the main function of the dictionary modeled. The prescriptive function of the dictionary defines its dominant typological characteristic. Thus, the application of the mathematical statistics method provided the terms selection for the prescriptive ecological thesaurus of civil aviation and defined basic typological characteristics of this dictionary.

КЛЮЧЕВЫЕ СЛОВА. Терминология, выборочная совокупность, генеральная совокупность, изоморфность, типологические параметры словаря.

KEY WORDS. Terminology, selection, entire assembly, isomorphism, typological characteristics.

Выработка научно обоснованных принципов отбора специальной лексики для отдельных типов терминологических словарей является одной из центральных задач современной терминографии [1; 217]. В настоящее время разработка терминологического массива с целью его лексикографического описания невозможна без применения методов математической статистики.

Цель данной статьи представить методику отбора терминов для формирования и обоснования терминологического массива нормативного двуязычного словаря-тезауруса терминов авиационной экологии. Мы покажем возможности применения методов математической статистики для выполнения задач прикладной лингвистики, а именно — прикладной терминографии и опишем использование критерия Пирсона хи-квадрат в разработке композиционных параметров нормативного двуязычного словаря-тезауруса терминов авиационной экологии.

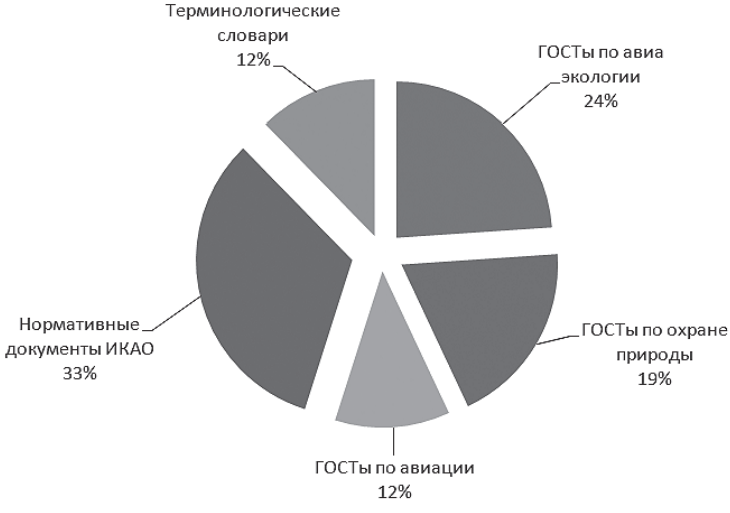
Отбор терминов авиационной экологии начинается с определения современного состояния данной предметной области в целом, а также актуальных экологических проблем гражданской авиации (ГА) в частности. Определив эти проблемы, возможно очертить круг источников терминов авиационной экологии, регламентирующих экологические нормы данной области.

К числу главных факторов, определяющих экологическое воздействие гражданской авиации на окружающую среду, относятся следующие: шум в районе аэропорта при взлете и посадке, акустическое воздействие на окружающую среду при полете по маршруту, выбросы загрязняющих веществ в районе аэропортов, выбросы парниковых газов в атмосферу в крейсерском полете, электромагнитное воздействие на окружающую среду, загрязнения, связанные с производством, утилизацией авиационных двигателей, топлив, масел и специальных жидкостей [2; 101], [3; 103].

При отборе источников терминологии авиационной экологии мы учитывали выделенные выше актуальные экологические проблемы, а также две возможные сферы применения любой отраслевой терминологии. Это сфера фиксации, представленная сборниками рекомендуемых терминов, государственными стандартами и специальными терминологическими словарями, и сфера функциониро-

вания, в которой термины существуют в текстах различных жанрово-стилистических разновидностей языка науки (статьи, обзоры, доклады и др.) [4; 158]. Так, источниками выборки для формирования терминологического массива по экологии гражданской авиации послужили следующие виды нормативных текстов: ГОСТы по авиационной экологии (10), ГОСТы по охране природы (6), ГОСТы по авиации (2), нормативные документы по экологии Международной организации гражданской авиации ИКАО (10), терминологические словари (2). В результате отбора терминов из этих источников с учетом существующих экологических проблем в области ГА нами получено пять выборок терминов, общим количеством 723 лексические единицы (терминов и терминологических словосочетаний). Соотношение количества терминов для каждой из выборок наглядно можно представить в виде диаграммы 1:

Диаграмма 1



Для того, чтобы сформировать полноценный терминологический массив для дальнейшего представления в специальном словаре, необходимо доказать, что выделенные выборки терминов характеризуются определенной степенью изоморфности, т.е. принадлежат одной генеральной совокупности, а следовательно, могут быть представлены в специальном словаре данной отрасли знания. Осуществить поставленную задачу возможно, используя методы математической описательной статистики.

Перейдем к этапу статистического анализа выборочной совокупности терминов. Основными понятиями математической статистики являются выборка (или выборочная совокупность случайных величин) и генеральная совокупность [5; 51]. Под выборочной совокупностью случайных величин в нашем исследовании будем понимать полученную выборку из 723 терминов, в то время как генеральная совокупность будет определяться как множество всех терминов исследуемой предметной области. Понятия выборочной и генеральной совокупности являются основополагающими в математической статистике, так как основной принцип применения статистических методов состоит в выявлении

и анализе закономерностей, присущих объектам в выборке, с целью установления возможности и достоверности перенесения сделанных выводов на генеральную совокупность. Следовательно, применение методов математической статистики позволит нам обосновать терминологический массив путем подтверждения количественного и качественного сходства между выделенными нами пятью выборками терминов, составляющих описываемую выборочную совокупность, а также разработать определяющие композиционные параметры словаря авиационной экологии.

Решение прикладных задач методами математической статистики, как правило, начинается с формулирования статистической гипотезы, иными словами — с постановки предположения о распределении случайных величин, которое будет подтверждено или опровергнуто в результате статистической обработки [6; 17] В данном исследовании *статистическая гипотеза* заключается в следующем утверждении: полученная выборочная совокупность терминов, сформированная из пяти различных источников, характеризуется однородным количественным и качественным распределением терминов, а значит, является изоморфной и принадлежит одной генеральной совокупности.

Перед тем как непосредственно перейти к сравнению выборок, была сформулирована нуль-гипотеза. В описательной статистике при сравнении двух (или нескольких) выборок нуль-гипотеза (далее H_0) формулируется так: *между изучаемыми выборками нет различия или, иначе, различие между ними несущественно*. Иными словами, все наши расчеты будут направлены на то, чтобы прийти к заключению, верна ли нуль-гипотеза или от нее нужно отказаться, и в действительности между выборками имеется существенная разница, а значит, их нельзя считать принадлежащими одной генеральной совокупности.

Для подтверждения нуль-гипотезы мы попарно сравнивали пять выборок терминов, представленных выше в диаграмме 1. Мы сравнивали выборки дважды, так как нам необходимо подтвердить количественное и качественное сходство между ними, а следовательно, сравнить два вида показателей: количественных (выраженных в частотном распределении терминов в корпусе научных текстов) и качественных (выраженных в соотношении научных и технических терминов в выборках).

Первичный частотный анализ выборочной совокупности терминов был проведен с использованием стандартных программных продуктов Microsoft Excel и Adobe Reader 7.0 на корпусе научных текстов, представленных в Научном вестнике МГТУ ГА (Московский государственный технический университет гражданской авиации) за 2012-2013 годы общим объемом 2104 страницы. Первичный частотный анализ пяти имеющихся выборок показал, что характер исходных данных далек от так называемого «нормального распределения» по закону Гаусса, так как количество терминов с частотным показателем $F = 0$ во много раз превышает количество терминов с другими показателями частотности.

Для сравнения выборок терминов с учетом непараметрического распределения показателей мы использовали критерий Пирсона *хи-квадрат*, наиболее часто применяемый для определения меры различия между наблюдаемыми и предполагаемыми (теоретическими) значениями. Метод *хи-квадрат* считается универсальным и применяется во многих исследованиях, он пригоден для статистического анализа распределения численностей разнообразных количе-

ственных материалов, относящихся ко всем статистическим шкалам. Формула критерия *хи-квадрат* имеет вид:

$$x^2 = \Sigma\ (f_0 - f_e)^2 / f_e$$

где f_0— наблюдаемые численности; f_e — предполагаемые (теоретические) численности.

Как указывалось выше, для проверки нуль-гипотезы мы дважды последовательно сравнили 5 выборок попарно: первую со второй, первую с третьей, первую с четвертой, первую с пятой, вторую с третьей и т.д. и т.п. Для определения критерия *хи-квадрат* для каждой пары выборок была составлена расчетная таблица (всего таких таблиц получилось 20). Согласно результатам частотного анализа, все термины были условно разделены на две большие группы: термины, которые встречаются в анализируемом корпусе один и более раз, и термины, не встретившиеся там ни разу. В таблицы расчетов количественного сравнения выборок были внесены именно эти две категории терминов: с частотностью ≥1 и термины с частотностью = 0. Вот пример такой расчетной таблицы для сравнения выборок 1 и 2 (табл. 1). Количества, которые были бы получены при принятии нуль-гипотезы, заключены в скобки. В правом углу буквенное обозначение клетки, необходимое для расчетов.

Таблица 1

Схема расчета x^2 для количественного сравнения выборки 1 и выборки 2:

Выборка	**Термины с частотностью ≥1**	**Термины с частотностью = 0**	**Всего**	**Долевые отношения, %**
1	34 A (25,02)	139 B (147,9)	173 (173)	55,6
2	11 C (19,98)	127 D (118,1)	138 (138)	44,4
Всего	45	266	311	100

Для расчета по формуле нам необходимо получить разности по клеткам A, B, C, D между наблюдаемыми и предполагаемыми численностями. При этом знак разности несущественен:

A f_A = 34-25,02 = 8,98,
B f_B = 139-147,9 = 8,9,
C f_C = 11-19,98 = 8,98,
Д f_D= 127-118,1= 8,9.

Далее, подставив формулу x^2, получим:

$$x^2 = 8{,}98^2/25{,}02 + 8{,}9^2/147{,}9 + 8{,}98^2/19{,}98 + 8{,}9^2/118{,}1 = 8{,}43$$

$$x^2 = \mathbf{8{,}43}$$

Определив значения критерия *хи-квадрат* для данной пары выборок, мы воспользовались таблицей уровней значимости *хи-квадрат*, в которой указаны значения критерия, подтверждающие нуль-гипотезу с определенной вероятностью ошибки (от 0,1 до 0,99). Для того, чтобы подставить в таблицу критических значений подсчитанное нами значение x^2, необходимо получить число степеней свободы для данного исследования, которое определяется по формуле:

$$df = (k-1)(c-1) = (2-1) \times (2-1) = 1,$$

где k — число столбцов, с — число строк в таблице с анализируемым материалом.

Из таблицы уровней значимости для одной степени свободы для критерия хи-квадрат получим $x^2_{0,99}$ = 6,6. Следовательно, полученная величина (она не намного больше) вполне достаточна для отвержения H_0. Есть все основания для содержательного вывода о том, что *имеется разница между частотными показателями Выборки 1 и Выборки 2*.

Аналогичным способом был рассчитан критерий *хи-квадрат* для остальных пар выборок. При этом для качественного сравнения выборок мы рассматривали их с точки зрения соотношения научных и технических терминов. И именно эти показатели (количество научных и количество технических терминов) были внесены в расчетные таблицы в колонки 1 и 2. Прежде чем представить полученные результаты статистического описания выборочной совокупности терминов, рассмотрим некоторые признаки, по которым тот или иной термин был отнесен к категории *научный* или *технический* термин.

Наше исследование проводилось на материале научно-технической терминологии ГА. Мы классифицировали понятия, представленные в данной области знания, на научные и технические по следующим принципам:

— если термин выражает абстрактное понятие (свойство, состояния, класс) данной предметной области, то это *научный термин* [7], [8];

— если термин выражает понятие, соотносимое с материальным предметом действительности, реальным объектом данной предметной области, объектом, воспринимаемым органами чувств человека, то это *технический термин*.

Итак, полученные результаты количественного и качественного сравнения выборок можно представить в виде следующих таблиц:

Таблица 2

Результаты сравнения выборок по количественным показателям

Выборки	Распределение показателей частотности
В 1 - В 2	Неоднородны (x^2 = 8,43, $x^2_{0,99}$ = 6,6)
В 1 - В 3	Неоднородны (x^2 = 7, $x^2_{0,99}$ = 6,6)
В 1 - В 4	Однородны (x^2 = 0,71, $x^2_{0,6}$ = 0,7)
В 1 - В 5	Однородны (x^2 = 0,58, $x^2_{0,6}$ = 0,7)
В 2 - В 3	Однородны (x^2 = 0,065, $x^2_{0,2}$ = 0,064)
В 2 - В 4	Неоднородны (x^2 = 13,9, $x^2_{0,99}$ = 6,6)
В 2 - В 5	Однородны (x^2 = 3,34, $x^2_{0,95}$ = 3,8)
В 3 - В 4	Неоднородны (x^2 = 11, $x^2_{0,99}$ = 6,6)
В 3 - В 5	Однородны (x^2 = 3,3, $x^2_{0,95}$ = 3,8)
В 4 - В 5	Однородны (x^2 = 2,09, $x^2_{0,9}$ = 2,7)

Таблица 3

Результаты сравнения выборок по качественным показателям

Выборки	Распределение качественных показателей
В 1 - В 2	Неоднородны (x^2 = 28,67, $x^2_{0,99}$ = 6,6)
В 1 - В 3	Неоднородны (x^2 = 11,56, $x^2_{0,99}$ = 6,6)
В 1 - В 4	Неоднородны (x^2 = 18,16, $x^2_{0,99}$ = 6,6)
В 1 - В 5	Неоднородны (x^2 = 20,2, $x^2_{0,99}$ = 6,6)
В 2 - В 3	Однородны (x^2 = 2,19, $x^2_{0,9}$ = 2,7)
В 2 - В 4	Однородны (x^2 = 3,95, $x^2_{0,95}$ = 3,8)
В 2 - В 5	Однородны (x^2 = 0,0091, $x^2_{0,05}$ = 0,0039)
В 3 - В 4	Однородны (x^2 = 0,0751, $x^2_{0,2}$ = 0,064)
В 3 - В 5	Однородны (x^2 = 1,7, $x^2_{0,8}$ = 1,64)
В 4 - В 5	Однородны (x^2 = 2,81, $x^2_{0,9}$ = 2,7)

Таким образом, в 6 случаях из 10 мы признаем выборки однородными по распределению показателей частотности с вероятностью ошибки p≤0,05-0,8. В случае *B1 - B3* разница распределения показателей частотности несущественна. Только в 3 случаях из 10 мы имеем дело с существенным различием между выборками.

В 6 случаях из 10 выборки однородны по распределению качественных показателей, т.е. по соотношению научных и технических терминов. Такие результаты совпадают с результатами, полученными при количественном сравнении выборок и представленными в табл. 2. Однако в данном случае становится очевидным существенное отличие выборки 1 от остальных выборок терминов. Нам не удалось статистически подтвердить качественную изоморфность B1 с B2, B3, B4, B5. В то же время количественная изоморфность была доказана, а следовательно, мы не можем исключить термины B1 из нашей выборочной совокупности.

В результате статистического анализа выборок установили, что наша статистическая гипотеза подтвердилась. Выборочная совокупность терминов оказалась на 70% однородной по количественным показателям и на 60% изоморфной по качественным показателям, а значит, считается принадлежащей одной генеральной совокупности. При этом источниками для выборки терминов, как указывалось выше, послужил корпус текстов, представленный в рамках нормативной парадигмы исследуемой области знания в сфере функционирования и фиксации.

С помощью методов математической статистики мы доказали количественную и качественную изоморфность выборок, правомерность их систематизации и представления в одном терминологическом словаре. Отобранный терминологический массив находится в рамках нормативной парадигмы исследуемой предметной области и определяет нормативную функцию в качестве доминирующей функции моделируемого словаря. При этом словарная функция закладывается как основной композиционный типологический параметр. Таким образом, нормативный характер источников устанавливает границы терминосистемы и обуславливает нормативность терминологического массива, а следовательно, определяет нормативный тип словаря. В заключение подчеркнем, что в силу того, что все композиционные параметры терминологического словаря взаимообусловлены, нормативный параметр словаря детерминирует его систематизирующий характер и определяет перспективу пользователя [9], [10].

СПИСОК ЛИТЕРАТУРЫ

1. Гринев-Гриневич С.В. Терминоведение: учеб. пособ. для студ. высш. учеб. заведений. М.: Академия, 2008. 304 с.

2. Скрипниченко С.Ю., Масленникова Г.Е., Плешаков А.И. О комплексном показателе экологичности эксплуатации воздушного судна // Научный вестник МГТУ ГА. 2012. № 179. С. 101-104.

3. Карпин Б.Н., Матягина А.М., Николайкин Н.И. Некоторые аспекты разработки федеральных авиационных правил по охране окружающей среды в гражданской авиации // Научный вестник МГТУ ГА. 2009. № 144. Серия «Воздушное право». С. 100-107.

4. Даниленко В.П. Русская терминология. Опыт лингвистического описания. М.: Наука, 1977. 246 с.

5. Кремер Н.Ш. Теория вероятностей и математическая статистика. Учебник. М.: ЮНИТИ-ДАНА, 2001. 213 с.

6. Кибзун А.И., Горяинова Е.Р., Наумов А.В., Сиротин А.Н. Теория вероятностей и математическая статистика. Базовый курс с примерами и задачами: Учебное пособие. М.: Физматлит, 2002. 157 с.

7. Скороходько Э.Ф. Вопросы перевода английской технической литературы. Изд. 2-е. М.: Наука, 1963. 92 с.

8. Горский Д.П. Вопросы абстракции и образования понятия. М.: Изд-во АН СССР, 1961. 353 с.

9. Hartmann, R.R.K. Lexicography. Critical concepts. L.; N.Y., 2004.

10. Varantola, K. Linguistic Corpora (Databases) and the Compilation of Dictionaries // A practical Guide to Lexicography / P. van Sterkenburg. Amsterdam; Philadelphia, 2003.

REFERENCES

1. Grinev-Grinevich, S.V. *Terminovedenie: ucheb. posob. dlja stud. vyssh. ucheb. zavedenij* [Terminology: manual for students of universities and institutions]. Moscow, 2008. 304 p. (in Russian).

2. Skripnichenko, S.Ju., Maslennikova, G.E., Pleshakov, A.I. On the complex refractive environmental operation of the aircraft. *Nauchnyj vestnik MGTU GA — Scientific Bulletin MSTUCA*. 2012. Number 179. Pp. 101-104. (in Russian).

3. Karpin B.N. Matyagina A.M., Nikolaykin N.I. Some aspects of the development ofthe federal aviation regulations for environmental protection in civil aviation. *Nauchnyj vestnik MGTU GA — Scientific Bulletin MSTUCA*. 2009. № 144. Series “Air Law”. Pp. 100-107. (in Russian).

4. Danilenko, V.P. *Russkaja terminologija. Opyt lingvisticheskogo opisanija* [Russian terminology. Experience linguistic description]. Moscow: Nauka, 1977. 246 p. (in Russian).

5. Kremer, N.Sh. *Teorija verojatnostej i matematicheskaja statistika. Uchebnik* [The theory of probability and mathematical statistics. Textbook]. Moscow, 2001. 213 p. (in Russian).

6. Kibzun, A.I., Gorjainova, E.R., Naumov, A.V., Sirotin, A.N. *Teorija verojatnostej i matematicheskaja statistika. Bazovyj kurs s primerami i zadachami: Uchebnoe posobie* [The theory of probability and mathematical statistics. Base course with examples and sums: Manual]. Moscow, 2002. 157 p. (in Russian).

7. Skorohod’ko, Je.F. *Voprosy perevoda anglijskoj tehnicheskoj literatury. Izd. 2-e* [Issues of English technical literature texts translation. Edition 2]. Moscow, 1963. 92 p. (in Russian).

8. Gorskij, D.P. *Voprosy abstrakcii i obrazovanija ponjatija* [Questions of abstraction and concept formation]. Moscow, 1961. 353 p. (in Russian).

9. Hartmann, R.R.K. Lexicography. Critical concepts. L.; N.Y., 2004.

10. Varantola, K. Linguistic Corpora (Databases) and the Compilation of Dictionaries // A practical Guide to Lexicography / P. van Sterkenburg. Amsterdam; Philadelphia, 2003.

Автор публикации

Шевцова Анна Юрьевна — старший преподаватель кафедры иностранных языков и межкультурной профессиональной коммуникации гуманитарных направлений Института истории и политических наук Тюменского государственного университета

Author of the publication

Anna Yu. Shevtsova — Senior Lecturer, Department of Foreign Languages and Intercultural Professional Communication, Institute of History and Political Sciences, Tyumen State University