



Анализ взаимосвязи наименований и качества различных групп пищевой продукции

Г.Д. Щербаков, В.В. Бессонов, Э.С.-А. Шахвалиева

ФГБУН «Федеральный исследовательский центр питания, биотехнологии и безопасности пищи»,
Устьинский пр-д, д. 2/14, г. Москва, 109240, Российская Федерация

Резюме

Введение. Анализ наименований пищевых продуктов является важной задачей, направленной на решение двух проблем: определения взаимосвязи между результатами исследований и отдельными применяемыми словами и, как следствие, получения достоверной с точки зрения качества пищевых продуктов классификации внутри подгрупп.

Цель исследования: анализ взаимосвязи между наименованиями пищевых продуктов и их микро- и макро- и триентным составом.

Материалы и методы. В качестве объектов исследования были выбраны результаты исследований хлебобулочных изделий, мясных полуфабрикатов и питьевого молока, выполненные в рамках федерального проекта «Укрепление общественного здоровья» в 2020–2021 годах лабораториями Роспотребнадзора. Использовались такие методы лингвистического анализа, как токенизация, удаление шумовых слов, стемминг и анализ N-грамм.

Результаты. Для группы хлебобулочных изделий были выделены слова, а также их отдельные составляющие, которые позволили разделить образцы по ранее полученным группам с различным содержанием натрия, белка и жира. Для группы мясных полуфабрикатов, где разделение было проведено по показателям натрия и жира, также был получен список слов, позволяющих провести обратную классификацию по наименованиям. Для молока питьевого был получен отрицательный результат, продукты, в которых было выявлено низкое содержание кальция, никак не отличались от других групп по своим наименованиям и их частям.

Заключение. Анализ наименований пищевых продуктов продемонстрировал потенциальную возможность для классификации пищевых продуктов по их наименованиям с целью оценки их вероятного микро- и макронутриентного состава. Необходимо провести ряд дальнейших исследований, направленных на расширение перечня анализируемых групп продукции, особенно входящих в потребительскую корзину.

Ключевые слова: качество пищевых продуктов, наименования продукции, лингвистический анализ, база данных химического состава пищевых продуктов, цифровая нутрициология, классификация пищевых продуктов.

Для цитирования: Щербаков Г.Д., Бессонов В.В., Шахвалиева Э.С.-А. Анализ взаимосвязи наименований и качества различных групп пищевой продукции // Здоровье населения и среда обитания. 2023. Т. 31. № 12. С. 82–90. doi: 10.35627/2219-5238/2023-31-12-82-90

Analysis of the Relationship between Names and Quality of Various Groups of Food Products

Grigory D. Shcherbakov, Vladimir V. Bessonov, Elina S.-A. Shakhvaliyeva

Federal Research Center of Nutrition, Biotechnology and Food Safety,
2/14 Ustyinsky Driveway, Moscow, 109240, Russian Federation

Summary

Introduction: The analysis of food names is an important task aimed at solving two problems, namely, determining the relationship between research results and individual words used and, as a result, obtaining a reliable, from the point of view of food quality, classification within subgroups.

Objective: To analyze the relationship between the names of food products and their micro- and macronutrient composition.

Materials and methods: The research object was the whole aggregate of the results of testing bakery products, processed meat products, and milk carried out within the framework of the Federal Project on Public Health Promotion in 2020–2021 by laboratories of the Russian Federal Service for Surveillance on Consumer Rights Protection and Human Welfare (Rospotrebnadzor). We applied such linguistic analysis methods as tokenization, noise word removal, stemming, and N-gram analysis.

Results: For bakery products, we selected words and their parts enabling us to divide samples into previously obtained groups with different contents of sodium, protein, and fat. For processed meat products, classified by the sodium and fat contents, we also compiled a list of words allowing a reverse classification by name. For fluid milk, we obtained a negative result since the products with the established low calcium content did not differ from other groups in terms of names and their parts.

Conclusions: The analysis of food names has demonstrated the potential for classifying foods by their names in order to assess their likely micro- and macronutrient composition. It is necessary to conduct a number of further studies aimed at expanding the list of analyzed product groups, especially those included in the consumer basket.

Keywords: food quality, product names, linguistic analysis, food chemical composition database, digital nutrition, food classification.

For citation: Shcherbakov GD, Bessonov VV, Shakhvaliyeva ES-A. Analysis of the relationship between names and quality of various groups of food products. *Zdorov'e Naseleniya i Sreda Obitaniya*. 2023;31(12):82–90. (In Russ.) doi: 10.35627/2219-5238/2023-31-12-82-90

Введение. Базы данных химического состава пищевых продуктов (БД ХСПП) являются важным аналитическим инструментом для пищевой промышленности и исследований в области питания. Они представляют собой всеобъемлющий источник информации о составе пищевых продуктов, содержании нутриентов и других соответствующих данных. Использование таких баз данных становится все более важным в последние годы, поскольку число исследований о влиянии рациона на здоровье неуклонно растет.

БД ХСПП используют для определения содержания нутриентов в пищевых продуктах, включая макроэлементы (например, углеводы, белки и жиры), витамины, минеральные вещества и биологически активные соединения. Информация о содержании нутриентов в пищевых продуктах может быть использована для мониторинга питания и оценки потребления пищи, а также для предоставления рекомендаций по питанию [1]. БД ХСПП также используют для создания инструментов оценки рациона питания и составления таблиц химического состава пищевых продуктов. Кроме того, БД ХСПП используются в эпидемиологических исследованиях для изучения связи между питанием и здоровьем, оценки рисков развития алиментарно-зависимых заболеваний [2].

БД ХСПП применяют для оценки воздействия производства пищевых продуктов на окружающую среду, например, углеродного следа пищевых производств, а также методов ведения сельского хозяйства [3]. К преимуществам использования БД ХСПП относятся их точность, надежность и доступность [4].

БД ХСПП предоставляют точную и актуальную информацию о содержании в пищевых продуктах нутриентов, важных с точки зрения питания и здоровья. Они также являются надежными источниками информации. Кроме того, БД ХСПП легкодоступны, поскольку они открыто размещены в Интернете или опубликованы в печатном формате [5]. Однако существуют некоторые ограничения, связанные с использованием БД ХСПП [6]. Например, некоторые базы данных могут быть недостаточно полными, чтобы собрать всю необходимую информацию о пищевом продукте. Кроме того, некоторые базы данных могут не включать данные из разных стран или регионов, а также могут не содержать подробной информации о способах приготовления (рецептурах) того или иного пищевого продукта или блюда [7, 8]. Для того чтобы нивелировать данные ограничения, исследователями применяется ряд мер.

Важным инструментом являются меры контроля качества (МКК) данных. МКК представляют собой установленные руководящие принципы, которые помогают обеспечить точность данных, предоставляя рекомендации по допустимым уровням ошибок или неточностей, а также другие критерии, такие как полнота и предельная давность данных [9]. МКК можно применять к различным типам источников данных, включая ответы на опросы, результаты лабораторных исследований или онлайн-отчеты [10]. Кроме того, они помогают повысить надежность

и точность любых выводов, полученных на основе анализа данных, собранных из этих ресурсов [11].

Вторая мера, принимаемая исследователями при работе с БД ХСПП, заключается в ограничении включения определенных типов источников данных в их анализ. Это помогает обеспечить включение только данных из надежных источников, точность которых была проверена до внесения их в базу [12]. Кроме того, задание критериев включения, таких как временные рамки или географический район сбора данных, дополнительно повышает достоверность и надежность любых выводов, сделанных на основе этих ресурсов [13]. Такие ограничения также могут уменьшить систематическую ошибку, возникающую из-за расхождений между системами маркировки или моделями рационов (нутриентные профили рациона) в разных странах, которые в противном случае могли бы негативно повлиять на интерпретацию результатов исследований [14].

Третий подход, часто встречающийся при использовании БД ХСПП, включает ограничение доступа к данным на основе определенных критериев, связанных с целями анализа или потребностями в отчетности, такими как режимы питания, желаемые переменные и т. д. Например, при анализе моделей вегетарианского питания из анализа исключаются все продукты, содержащие продукты животного происхождения, тогда как при анализе всего населения такие элементы могут оставаться частью анализа в зависимости от конкретной поставленной цели [15]. Кроме того, некоторые нутриенты могут нуждаться в установлении границ на основе ограничений, например, для подбора продуктов для больных диабетом второго типа [16]. Добавление фильтров, таких как возраст, пол и др., позволяет ограничить перечень потенциально потребляемой продукции [17]. Наконец, установление исходных параметров заблаговременно значительно ограничивает разброс данных, что позволяет лучше сравнивать статистические показатели, используемые в процессе, тем самым повышая достоверность проводимых анализов более точных моделей [18].

Классификация пищевой продукции, в том числе и для использования при систематизации данных химического состава, оценке фактического потребления пищи и других исследованиях, связанных с питанием, представляет собой сложную задачу, а универсальный подход к ней еще не выработан.

Поэтому поиск инструментов и классификационных признаков, позволяющих выстроить систему организации данных в БД ХСПП, представляется крайне актуальным.

Цель исследования – провести лексический анализ наименований пищевых продуктов трех категорий для целей классификации и дальнейшего анализа.

Материалы и методы. В качестве материала для обработки и анализа была выбрана база результатов исследований качества и безопасности пищевых продуктов, выполненных в рамках федерального проекта «Укрепление общественного здоровья» национального проекта «Демография»¹

¹ Паспорт федерального проекта «Формирование системы мотивации граждан к здоровому образу жизни, включая здоровое питание и отказ от вредных привычек», реализующийся в рамках исполнения Указа Президента Российской Федерации от 07.05.2018 № 204 «О национальных целях и стратегических задачах развития Российской Федерации на период до 2024 года».

в 2020 и 2021 годах. Для анализа были отобраны результаты исследований химического состава хлебобулочных изделий, мясных полуфабрикатов и питьевого молока.

Для анализа наименований использовали группу основных методов лингвистического анализа. Токенизация – процесс обработки естественного языка, который включает разбиение необработанного текста на более мелкие фрагменты, называемые токенами. В данной работе в качестве токенов рассматривали отдельные слова. В качестве шумовых использовали слова, внесенные в пакет tidytext, такие как «и», «или», «из», «с» и подобные. Всего массив слов для исключения составлял 159 различных единиц. Стемминг – сокращение слова до его грамматической основы, неизменяемой при склонении. Для стемминга использовали алгоритм Snowball («Снежок») [19]. Анализ N-грамм – анализ последовательностей из N слов. В данной работе рассматривали только биграммы – последовательности из двух слов, при этом варианты перестановки слов местами рассматривались как одна форма.

В данной статье рассматривались результаты кластерного анализа указанных групп продуктов, полученные авторами ранее. Кластеризация проводилась методом K-средних, оптимальное число кластеров определялось на основании 30 показателей, таких как индекс Хьюберта и индекс Фридмана. Показатели для кластеризации выбирали на основании анализа их распределения с максимально возможным включением критически значимых нутриентов. Для целей настоящего анализа в качестве критически значимых нутриентов рассматривали содержание белка, жира, пищевой соли. Для оценки содержания в образцах пищевых продуктов соли использовали данные о количественном содержании в них натрия.

Алгоритм и соответствующие вычисления реализованы на языке R версии 4.2.1 в среде разработки RStudio [20].

Результаты. Был проведен анализ наиболее часто встречающихся слов в образцах хлебобулочных изделий. По результатам кластеризации [21] было выделено три группы хлеба с различным содержанием критически значимых нутриентов. Первый кластер (301 образец) характеризовался наименьшим содержанием белка и наибольшим содержанием натрия, второй кластер (194 образца) – наибольшим содержанием жира и средним содержанием натрия, третий кластер (306 образцов) – наибольшим содержанием белка и наименьшим содержанием натрия.

После подобного разделения интерес представляло проведение обратного анализа, то есть было необходимо выделить наименования образцов или их части, которые наиболее часто встречались в соответствующих кластерах (табл. 1). Необходимо отметить, что даже без применения стемминга видны слова, которые по частоте встречаемости позволяли дифференцировать одну группу от другой. Так, практически все образцы, содержащие в себе слово «батон», относились к третьему и частично ко второму кластеру. А практически все образцы хлеба

«Дарницкого» относились к первому кластеру, то есть к группе с высоким содержанием натрия.

После стемминга можно было сделать дополнительные выводы о том, что во втором кластере продуктов чаще встречается указание на муку и ее сорт, в том числе чаще – указание именно на муку первого или высшего сорта (табл. 2). Это может свидетельствовать о том, что данные категории хлеба специально изготавливаются с большим содержанием жира для улучшения органолептических свойств и, как следствие, потенциальным установлением у потребителя связи между сортностью и вкусом.

Для подтверждения полученных выводов также были проанализированы биграммы, входящие в наименования образцов (табл. 3). В данном случае не были исключены стоп-слова для более корректного анализа. Были подтверждены первичные предположения о том, что существует дополнительная возможность для эмпирического отнесения хлеба к одному или другому кластеру из полученных.

Далее был проведен анализ наиболее часто встречающихся слов на этикетках образцов мясных полуфабрикатов и колбасных изделий. По результатам кластеризации [22] было выделено три группы мясных полуфабрикатов с различным содержанием жира и пищевой соли (натрия). В первый кластер (457 образцов) попали образцы

Таблица 1. Частота встречаемости слов в наименованиях хлеба

Table 1. Frequency of occurrence of specific words in the names of bread

Слово / Word	Частота / Frequency		
	Первый кластер / First cluster	Второй кластер / Second cluster	Третий кластер / Third cluster
хлеб [bread]	293	81	246
упаковке [packed]	99	78	129
муки [flour]	67	53	127
формовой [tin]	100	12	80
пшеничный [wheat]	64	27	93
пшеничной [wheat]	47	45	86
батон [loaf]	10	107	49
сорта [grade]	37	33	87
нарезанный [sliced]	39	57	45
нарезной [threaded]	8	65	19
изделие [product]	22	21	27
дарницкий [Darnitsky]	58	2	8
ржаной [rye]	36	7	24
высшего [premium]	11	17	35
ржано [rye]	40	7	15
хлебопекарной [for baking bread]	21	13	28
подовый [hearth]	26	8	25
хлебобулочное [bakery]	20	15	20
первого [first]	13	10	31
изделия [product]	19	7	25

Примечание к табл. 1–9. В данной работе рассматриваются образцы отечественной пищевой продукции, поэтому перевод слов, стемм и биграмм на английский язык в таблицах является условным.

Note to Tables 1–9: Samples of Russian food products are examined in this work. The translation of words, stems and bigrams into English provided in the tables is therefore conditional.

Таблица 2. Частота встречаемости стемм в наименованиях хлеба
Table 2. Frequency of occurrence of stems in the names of bread

Стемма / Stem	Частота / Frequency		
	Первый кластер / First cluster	Второй кластер / Second cluster	Третий кластер / Third cluster
хлеб [bread]	293	81	246
пшеничн [wheat]	122	75	185
упаковк [packed]	101	80	132
мук [flour]	70	53	127
формов [tin]	105	15	81
нареза [cut]	54	66	60
сорт [grade]	42	38	99
батон [loaf]	12	111	52
ржан [rye]	78	14	40
издел [product]	41	28	52
нарезн [cut]	13	67	19
хлебобулочн [bakery]	24	17	31
высш [premium]	11	17	40
дарницк [Darnitsky]	58	2	8
хлебопекарн [bakery]	23	13	29
подов [hearth]	27	10	27
перв [first]	14	10	33
бел [white]	8	2	31
нов [new]	22	6	13
смес [mixed]	23	3	14

Таблица 3. Частота встречаемости биграмм в наименованиях хлеба
Table 3. Frequency of occurrence of bigrams in the names of bread

Биграмма / Bigram	Частота / Frequency		
	Первый кластер / First cluster	Второй кластер / Second cluster	Третий кластер / Third cluster
в упаковке [packed]	98	76	122
из пшеничной [from wheat]	19	37	56
пшеничной муки [wheat flour]	24	31	47
хлеб пшеничный [wheat bread]	24	10	66
нарезанный в [cut in]	25	38	34
из муки [from wheat]	18	11	51
формовой в [tin in]	42	2	30
батон нарезной [threaded loaf]	4	55	15
хлеб дарницкий [Darnitsky bread]	56	2	6
высшего сорта [premium grade]	9	16	34
ржано пшеничный [rye & wheat]	36	7	15
первого сорта [first grade]	13	9	30
муки высшего [premium flour]	8	16	27
изделие хлебобулочное [bakery product]	16	15	16
сорта формовой [grade tin]	14	1	31
пшеничный из [wheat from]	9	2	33
пшеничной хлебопекарной [wheat for baking bread]	11	9	23
хлебопекарной муки [bread flour]	13	10	18
из смеси [of a mix]	20	3	14
муки первого [flour of first]	7	6	24

с минимальным содержанием жира и максимальным содержанием натрия, во вторую группу (179 образцов) – со средним содержанием жира и минимальным содержанием натрия, а в третью (288 образцов) – с высоким содержанием натрия и максимальным содержанием жира.

Для третьей группы продуктов характерно то, что среди их наименований встречается довольно большое количество слов с высокой частотой повторения, что требует более детального анализа, для первичного представления были взяты 20 наиболее часто встречающихся слов (табл. 4).

Необходимо отметить, что именно докторская колбаса наиболее часто относилась к первому кластеру. Однако остальные виды мясных полуфабрикатов распределялись равномерно, с учетом размеров соответствующих групп, среди кластеров.

После стемминга (табл. 5) также невозможно было отделить какие-то конкретные наименования, которые характеризовали бы тот или иной кластер.

Также были проанализированы биграммы, входящие в наименования образцов (табл. 6). В данном случае для более корректного анализа стоп-слова не были исключены. Интерес представляло наличие указания на категорию и вид продукта, именно как применяемый термин.

Можно сделать вывод о том, что продукты категории «Б» чаще встречались в первом и в третьем кластере, чем во втором. Обратную ситуацию наблюдали с продуктами из мяса птицы, что соответствует логике, что данные изделия являются менее жирными. Данный вывод также поддерживает то, что практически все образцы докторской колбасы также были отнесены к первому кластеру.

Таблица 4. Частота встречаемости слов в наименованиях мясных полуфабрикатов
Table 4. Frequency of occurrence of specific words in the names of processed meat products

Слово / Word	Частота / Frequency		
	Первый кластер / First cluster	Второй кластер / Second cluster	Третий кластер / Third cluster
колбасное [sausage]	277	96	164
изделие [product]	271	93	168
категории [category]	222	91	161
продукт [product]	222	86	149
вареное [boiled]	227	82	147
мясной [meat]	198	74	141
сосиски [sausages]	147	42	84
сардельки [sausages]	122	63	71
колбаса [sausage]	131	33	52
вареные [boiled]	83	41	49
изделия [products]	85	31	44
колбасные [sausage]	79	31	44
шпикачки [pork sausage]	30	28	70
вареная [boiled]	67	17	35
мясное [meat]	58	20	35
мяса [meat]	70	21	20
охлажденные [chilled]	51	29	29
птицы [poultry]	59	19	17
молочные [milk]	32	14	41
докторская [doctor's]	62	15	8

Таблица 5. Частота встречаемости стемм в наименованиях мясных полуфабрикатов
Table 5. Frequency of occurrence of stems in the names of processed meat products

Стемма / Stem	Частота / Frequency		
	Первый кластер / First cluster	Второй кластер / Second cluster	Третий кластер / Third cluster
варен [boiled]	398	142	239
издел [product]	358	124	213
колбасн [sausage]	356	127	209
мясн [meat]	284	99	192
категор [category]	233	97	168
продукт [product]	223	86	151
сосиск [sausage]	147	42	84
охлажден [chilled]	126	55	88
сардельк [sausage]	122	63	71
колбас [sausage]	135	34	54
шпикачк [pork sausage]	30	28	70
мяс [meat]	73	21	20
молочн [milk]	49	19	45
докторск [doctor's]	70	19	13
птиц [poultry]	59	19	17
сорт [grade]	46	6	22
мясодержа [containing meat]	30	14	14
сочн [juicy]	15	16	15
сливочн [creamy]	20	5	18

Таблица 6. Частота встречаемости биграмм в наименованиях мясных полуфабрикатов
Table 6. Frequency of occurrence of bigrams in the names of processed meat products

Биграмма / Bigram	Частота / Frequency		
	Первый кластер / First cluster	Второй кластер / Second cluster	Третий кластер / Third cluster
колбасное изделие [sausage product]	188	69	114
мясной продукт [meat product]	173	60	129
изделие вареное [boiled product]	135	57	85
категории б [of category b]	106	44	106
продукт категории [product of category]	87	35	67
изделие колбасное [sausage product]	75	23	45
колбасное вареное [boiled sausage]	73	21	43
вареное категории [boiled of category]	49	19	44
из мяса [of meat]	70	21	20
категории в [of category c]	52	24	34
колбаса вареная [boiled sausage]	63	15	30
мяса птицы [poultry]	58	19	17
колбасные изделия [sausage products]	46	14	26
категории а [of category a]	51	17	13

Был проведен анализ наиболее часто встречающихся слов в образцах молока. По результатам кластеризации [23] были получены три группы: первая группа (49 образцов) – молоко с содержанием жира как 3,2, так и 2,5 %, но с содержанием кальция от 500 до 700 мг/кг, вторая группа (129 образцов) – молоко с содержанием жира 3,2 % и содержанием кальция 1050–1200 мг/кг, третья группа (112 образцов) – молоко с содержанием жира 2,5 % и содержанием кальция 1050–1200 мг/кг.

Очевидно, что в названии фигурирует указание жирности молока, поэтому видно разделение по соответствующим кластерам, а также выявлены образцы, которые по результатам фактических измерений не соответствуют заявленному на этикетке (табл. 7).

После стемминга (табл. 8) были получены аналогичные результаты, отсутствовали специфические формы наименований, которые указывали бы на возможность выделить первый кластер среди остальных.

Таблица 7. Частота встречаемости слов в наименованиях молока питьевого
Table 7. Frequency of occurrence of specific words in the names of milk

Слово / Word	Частота / Frequency		
	Первый кластер / First cluster	Второй кластер / Second cluster	Третий кластер / Third cluster
молоко [milk]	45	138	119
питьевое [fluid]	43	125	103
пастеризованное [pasteurized]	27	62	59
2,5	24	5	100
ультрапастеризованное [ultra-pasteurized]	17	56	50
м.д.ж. [weight fraction of fat]	19	55	48
3,2	12	101	5
жира [fat]	19	51	48
массовой [weight]	8	22	32
долей [fraction]	8	21	32
массовая [weight]	11	32	16
доля [fraction]	10	32	16

Таблица 8. Частота встречаемости стемм в наименованиях молока питьевого
Table 8. Frequency of occurrence of stems in the names of milk

Стемма / Stem	Частота / Frequency		
	Первый кластер / First cluster	Второй кластер / Second cluster	Третий кластер / Third cluster
молок [milk]	45	140	121
питьев [fluid]	43	125	105
пастеризова [pasteurized]	27	66	59
2,5	24	5	100
ультрапастеризова [ultra-pasteurized]	17	56	50
м.д.ж. [weight fraction of fat]	19	55	48
дол [fraction]	19	54	48
массов [weight]	19	54	48
3,2	12	101	5
жир [fat]	19	51	48

Таблица 9. Частота встречаемости биграмм в наименованиях молока питьевого
Table 9. Frequency of occurrence of bigrams in the names of milk

Биграмма / Bigram	Частота / Frequency		
	Первый кластер / First cluster	Второй кластер / Second cluster	Третий кластер / Third cluster
молоко питьевого [fluid milk]	37	118	98
питьевое пастеризованное [potable pasteurized]	24	55	50
питьевое ультрапастеризованное [potable ultra-pasteurized]	16	54	46
массовой долей [weight fraction]	8	21	32
долей жира [fraction of fat]	8	20	32

Также были проанализированы биграммы, входящие в наименования образцов (табл. 9). В данном случае также для более корректного анализа стоп-слова не были исключены. Были подтверждены первичные результаты о том, что отсутствует дополнительная возможность для эмпирического отнесения молока к первому кластеру с меньшим содержанием кальция.

Обсуждение. Классификация пищевых продуктов – это процесс их распределения по категориям на основе их сходства. Это важный шаг в разработке соответствующих баз данных, поскольку он позволяет эффективно находить и сравнивать пищевую ценность продуктов между собой. В литературе представлены различные подходы к классификации пищевых продуктов и описаны преимущества и недостатки каждого из них.

Один из подходов к классификации пищевых продуктов (КПП) основан на их составе. Этот подход привлекателен тем, что позволяет сравнивать продукты на основе их нутриентного профиля, выделяя схожие продукты и их виды. Этот подход также может быть полезен при поиске продуктов, отвечающих определенным диетическим требованиям, таким как низкоуглеводные или веганские рационы [24]. Было отмечено, что этот подход может неточно отражать истинный состав пищевого продукта, поскольку некоторые пищевые вещества могут быть не включены в анализ [25].

Подходы на основе лингвистического анализа используют методы обработки естественного языка для определения релевантных ключевых слов, характеризующих блюда, представленные в онлайн-рецептах или меню ресторанов, чтобы классифицировать их в соответствии с местом их происхождения или типами кухни [26]. Эти методы в значительной степени зависят от точности описания блюд, которые могут быть затруднены из-за тонких различий между кухнями даже в пределах одной группы, например итальянской и итало-американской кухни. Кроме того, они не учитывают содержание нутриентов, которое также может помочь отличить конкретные блюда, принадлежащие к определенной кухне, поэтому необходимо провести

дополнительные исследования, прежде чем этот подход станет широко использоваться.

Визуальные особенности, извлеченные из изображений с использованием методов компьютерного зрения, также недавно были исследованы с многообещающими результатами с точки зрения точности и масштабируемости [27]. Преимущество этого подхода заключается в том, что он более понятен, чем текстовые методы, поскольку визуальные характеристики часто теснее связаны с региональной кухней, чем просто слова, например, японские суши по сравнению с американскими суши-роллами. Однако для извлечения визуальных признаков требуются специализированные модели глубокого обучения, построенные на больших наборах данных, что может ограничивать его использование небольшими организациями с меньшим количеством вычислительных ресурсов, чем необходимо для реализации подобных моделей.

Напротив, модели, основанные на паттернах использования ингредиентов, продемонстрировали потенциал в качестве альтернативного метода, способного точно классифицировать различные блюда в разных культурах без необходимости большого объема данных [28]. Анализируя онлайн-рецепты, можно определить частые сочетания ингредиентов, связанные с конкретными типами блюд, а также культурные предпочтения, что позволяет проводить точную классификацию без прямого доступа к наборам данных изображений блюд. Хотя этот метод по-прежнему в значительной степени зависит от ручных процессов описания ингредиентов, он потенциально может обеспечить эффективный способ классификации ранее неизвестных блюд, как только будет предоставлено достаточно обучающих примеров.

Учитывая вышеизложенное, можно сделать вывод о том, что современный потребитель является внимательным, изучающим доступную ему информацию о продукте и интересующимся потенциальным влиянием продукта на состояние своего здоровья.

Таким образом, государственная политика в области здоровьесбережения должна быть направлена на повышение возможности осуществления корректного выбора пищевых продуктов (осознанный выбор для формирования здорового рациона) посредством управления как доступными торговыми наименованиями, которые будут достоверно отражать состав или категорию продукта, так и максимально полным составом его маркировки.

В российском законодательстве требования к наименованиям продуктов установлены в ТР ТС 022², а именно наименование пищевой продукции, указываемое в маркировке, должно позволять относить продукцию к пищевой продукции, достоверно ее характеризовать и позволять отличать ее от другой пищевой продукции. Однако данная формулировка ограничена тем, что наименования должны быть сформулированы в соответствии с действующими техническими регламентами. Таким образом, продукция не обязательно будет иметь

² ТР ТС 022/2011 «Технический регламент Таможенного союза. Пищевая продукция в части ее маркировки». Введен в действие решением Комиссии Таможенного союза от 09.12.2011 № 881 «О принятии технического регламента Таможенного союза «Пищевая продукция в части ее маркировки»».

наименования, которые могли бы позволить отнести ее к тому или иному виду, предусмотренному национальными или международными стандартами. Это было бы необходимым условием для соотнесения показателей качества и соответствующих наименований продуктов.

При анализе наименований различных групп пищевых продуктов было выявлено наличие взаимосвязи между наличием отдельных слов и словосочетаний и химическим составом этих продуктов. В частности, исследование показало, что данная связь имеется у тех групп продуктов, где наименования являются более описательными, то есть для хлебобулочных изделий и для мясных полуфабрикатов и колбасных изделий. Однако эта связь должна быть доведена до потребителя, так как само по себе включение таких конкретных обозначений, как хлеб «Дарницкий» или колбаса «Докторская», в наименование продукта не ассоциируется с тем или иным уровнем содержания нутриента. Для молока же подобная связь отсутствует, то есть отсутствует какая-либо возможность разделения молока с низким и средним содержанием кальция. Для потребителя это означает невозможность самостоятельной оценки уровня потребляемого кальция, что особенно критично для детей и пожилых людей.

Учитывая, что стандарты на соответствующие группы продуктов являются определяющими в части формирования наименований, данное исследование демонстрирует важность стандартизации как деятельности по косвенному управлению ожиданиями потребителей пищевой продукции. С точки зрения государственной задачи по управлению качеством продуктов, необходимо развивать стандарты вида «общие технические условия» для соответствующих групп продукции. Закрепление сортности не ограничит торговые возможности, но повысит прозрачность выбора для покупателей.

Заключение. Данное исследование показало, что для некоторых групп продуктов наименования или их отдельные части могут быть ассоциированы с химическим составом этих продуктов. Данный факт может быть использован в первую очередь для систематизации данных химического состава пищевых продуктов в БД ХСПП, создания классификаций пищевой продукции, дальнейших исследований как для оценки восприятия потребителем продукции, так и для составления более корректных моделей потребления, учитывающих специфику наличия выявленных подгрупп в различных регионах. Будущие исследования должны быть сосредоточены на изучении более конкретных аспектов этой взаимосвязи, например региональных особенностей и различных трактовок при переводе наименований импортной продукции. Такие исследования могли бы в итоге дать возможность разработки дополнительной классификации пищевых продуктов для создания достоверной базы данных химического состава пищевых продуктов.

СПИСОК ЛИТЕРАТУРЫ

1. Drewnowski A, Rehm CD. Consumption of added sugars among US children and adults by food purchase location and food source. *Am J Clin Nutr.* 2014;100(3):901–907. doi: 10.3945/ajcn.114.089458
2. Thompson FE, Subar AF, Loria CM, Reedy JL, Baranowski T. Need for technological innovation in dietary assessment. *J Am Diet Assoc.* 2010;110(1):48–51. doi: 10.1016/j.jada.2009.10.008
3. Costa Tpd, Gillespie J, Pelc K, et al. Life cycle assessment tool for food supply chain environmental evaluation. *Sustainability.* 2023;15(1):718. doi: 10.3390/su15010718
4. Delgado A, Issaoui M, Vieira MC, Saraiva de Carvalho I, Fardet A. Food composition databases: Does it matter to human health? *Nutrients.* 2021;13(8):2816. doi: 10.3390/nu13082816
5. Elmadfa I, Meyer AL. Importance of food composition data to nutrition and public health. *Eur J Clin Nutr.* 2010;64(Suppl 3):S4–S7. doi: 10.1038/ejcn.2010.202
6. Vrontzos G, Pardalos P. Data mining and optimisation issues in the food industry. *Int J Sustain Agric Manag Inform.* 2017;3(1):44–64. doi: 10.1504/IJSAMI.2017.10003796
7. Vonthron S, Perrin C, Soulard CT. Foodscape: A scoping review and a research agenda for food security-related studies. *PLoS One.* 2020;15(5):e0233218. doi: 10.1371/journal.pone.0233218
8. Zhang D, Ouyang S, Cai M, et al. FADB–China: A molecular-level food adulteration database in China based on molecular fingerprints and similarity algorithms prediction expansion. *Food Chem.* 2020;327:127010. doi: 10.1016/j.foodchem.2020.127010
9. Wang Z, Talburt JR, Wu N, Dagtas S, Zozus MN. A rule-based data quality assessment system for electronic health record data. *Appl Clin Inform.* 2020;11(4):622–634. doi: 10.1055/s-0040-1715567
10. Marshall NW, Mackenzie A, Honey ID. Quality control measurements for digital x-ray detectors. *Phys Med Biol.* 2011;56(4):979–999. doi: 10.1088/0031-9155/56/4/007
11. Bietenbeck A, Thaler MA, Lupp PB, Klawonn F. Stronger together: Aggregated Z-values of traditional quality control measurements and patient medians improve detection of biases. *Clin Chem.* 2017;63(8):1377–1387. doi: 10.1373/clinchem.2016.269845
12. Eržen N, Rayner M, Pravst I. A comparative evaluation of the use of a food composition database and nutrition declarations for nutrient profiling. *J Food Nutr Res.* 2015;54(2):93–100. doi: 10.13140/2.1.2096.0000
13. Wang H, Herforth AW, Xi B, Zou Z. Validation of the diet quality questionnaire in Chinese children and adolescents and relationship with pediatric overweight and obesity. *Nutrients.* 2022;14(17):3551. doi: 10.3390/nu14173551
14. Shamim K, Khan SA, Ahmad S. Consumers' understanding of nutrition labels for ultra-processed food products. *J Public Aff.* 2022;22(1):e2398. doi: 10.1002/pa.2398
15. Jones NRV, Forouhi NG, Khaw KT, Wareham NJ, Monsivais P. Accordance to the dietary approaches to stop hypertension diet pattern and cardiovascular disease in a British, population-based cohort. *Eur J Epidemiol.* 2018;33(2):235–244. doi: 10.1007/s10654-017-0354-8
16. Cucuzzella M, Riley K, Isaacs D. Adapting medication for type 2 diabetes to a low carbohydrate diet. *Front Nutr.* 2021;8:688540. doi: 10.3389/fnut.2021.688540
17. Bennett G, Bardon LA, Gibney ER. A comparison of dietary patterns and factors influencing food choice among ethnic groups living in one locality: A systematic review. *Nutrients.* 2022;14(5):941. doi: 10.3390/nu14050941
18. Kirk D, Catal C, Tekinerdogan B. Precision nutrition: A systematic literature review. *Comput Biol Med.* 2021;133:104365. doi: 10.1016/j.compbiomed.2021.104365
19. Bounabi M, El Moutaouakil K, Satori K. A comparison of text classification methods using different stemming techniques. *Int J Comput Appl Technol.* 2019;60(4):298–306. doi: 10.1504/IJCAT.2019.101171

<https://doi.org/10.35627/2219-5238/2023-31-12-82-90>
Original Research Article

20. Silge J, Robinson D. tidytext: Text mining and analysis using tidy data principles in R. *JOSS*. 2016;1(3):37. doi: 10.21105/joss.00037
21. Щербаков Г.Д., Бессонов В.В. Подходы к алгоритму анализа результатов исследований микро- и макронутриентного состава хлебобулочных изделий. Сообщение второе // Здоровье населения и среда обитания. 2023;31(1):53-59. <https://doi.org/10.35627/10.35627/2219-5238/2023-31-1-53-59>
22. Щербаков Г.Д. Применение метода Монте-Карло для моделирования поступления различных нутриентов на примере мясных полуфабрикатов Основы здорового питания и пути профилактики алиментарно-зависимых заболеваний : Материалы V Школы молодых ученых с международным участием, Москва, 09–11 ноября 2022 года. – Москва: Федеральное государственное бюджетное учреждение науки Федеральный исследовательский центр питания, биотехнологии и безопасности пищи, 2022. С. 121-122.
23. Щербаков Г.Д., Бессонов Г.В. Алгоритм обработки и анализа результатов исследований микро- и макронутриентного состава молока // Здоровье населения и среда обитания. 2022;(8):64-72. <https://doi.org/10.35627/2219-5238/2022-30-8-64-72>
23. Щербаков Г.Д., Бессонов Г.В. Алгоритм обработки и анализа результатов исследований микро- и макронутриентного состава молока // Здоровье населения и среда обитания. 2022;(8):64-72. <https://doi.org/10.35627/2219-5238/2022-30-8-64-72>
24. Akbay A, Elhan A, Ozcan C, Demirtaş S. Hierarchical cluster analysis as an approach for systematic grouping of diet constituents on basis of fatty acid, energy and cholesterol content: Application on consumable lamb products. *Med Hypotheses*. 2000;55(2):147-154. doi: 10.1054/mehy.1999.1038
25. Ortenzi F, Kolby M, Lawrence M, et al. Limitations of the food compass nutrient profiling system. *J Nutr*. 2023;153(3):610-614. doi: 10.1016/j.tjnut.2023.01.027
26. Kicherer H, Dittrich M, Grebe L, Scheible C, Klinger R. What you use, not what you do: Automatic classification and similarity detection of recipes. *DKE*. 2018;117:252-263. doi: 10.1016/j.datak.2018.04.004
27. Subhi MA, Ali SH, Mohammed MA. Vision-based approaches for automatic food recognition and dietary assessment: A survey. *IEEE Access*. 2019;7:35370-35381. doi: 10.1109/ACCESS.2019.2904519
28. He H, Kong F, Tan J. DietCam: Multiview food recognition using a multikernel SVM. *IEEE J Biomed Health Inform*. 2016;20(3):848-855. doi: 10.1109/JBHI.2015.2419251
29. Vrontzos G, Pardalos P. Data mining and optimisation issues in the food industry. *Int J Sustain Agric Manag Inform*. 2017;3(1):44-64. doi: 10.1504/IJSAMI.2017.10003796
30. Vonthron S, Perrin C, Soulard CT. Foodscape: A scoping review and a research agenda for food security-related studies. *PLoS One*. 2020;15(5):e0233218. doi: 10.1371/journal.pone.0233218
31. Zhang D, Ouyang S, Cai M, et al. FADB–China: A molecular-level food adulteration database in China based on molecular fingerprints and similarity algorithms prediction expansion. *Food Chem*. 2020;327:127010. doi: 10.1016/j.foodchem.2020.127010
32. Wang Z, Talburt JR, Wu N, Dagtas S, Zozus MN. A rule-based data quality assessment system for electronic health record data. *Appl Clin Inform*. 2020;11(4):622-634. doi: 10.1055/s-0040-1715567
33. Marshall NW, Mackenzie A, Honey ID. Quality control measurements for digital x-ray detectors. *Phys Med Biol*. 2011;56(4):979-999. doi: 10.1088/0031-9155/56/4/007
34. Bietenbeck A, Thaler MA, Lupp PB, Klawonn F. Stronger together: Aggregated Z-values of traditional quality control measurements and patient medians improve detection of biases. *Clin Chem*. 2017;63(8):1377-1387. doi: 10.1373/clinchem.2016.269845
35. Erzen N, Rayner M, Pravst I. A comparative evaluation of the use of a food composition database and nutrition declarations for nutrient profiling. *J Food Nutr Res*. 2015;54(2):93-100. doi: 10.13140/2.1.2096.0000
36. Wang H, Herforth AW, Xi B, Zou Z. Validation of the diet quality questionnaire in Chinese children and adolescents and relationship with pediatric overweight and obesity. *Nutrients*. 2022;14(17):3551. doi: 10.3390/nu14173551
37. Shamim K, Khan SA, Ahmad S. Consumers' understanding of nutrition labels for ultra-processed food products. *J Public Aff*. 2022;22(1):e2398. doi: 10.1002/pa.2398
38. Jones NRV, Forouhi NG, Khaw KT, Wareham NJ, Monsivais P. Accordance to the dietary approaches to stop hypertension diet pattern and cardiovascular disease in a British, population-based cohort. *Eur J Epidemiol*. 2018;33(2):235-244. doi: 10.1007/s10654-017-0354-8
39. Cucuzzella M, Riley K, Isaacs D. Adapting medication for type 2 diabetes to a low carbohydrate diet. *Front Nutr*. 2021;8:688540. doi: 10.3389/fnut.2021.688540
40. Bennett G, Bardon LA, Gibney ER. A comparison of dietary patterns and factors influencing food choice among ethnic groups living in one locality: A systematic review. *Nutrients*. 2022;14(5):941. doi: 10.3390/nu14050941
41. Kirk D, Catal C, Tekinerdogan B. Precision nutrition: A systematic literature review. *Comput Biol Med*. 2021;133:104365. doi: 10.1016/j.compbimed.2021.104365
42. Bounabi M, El Moutaouakil K, Satori K. A comparison of text classification methods using different stemming techniques. *Int J Comput Appl Technol*. 2019;60(4):298-306. doi: 10.1504/IJCAT.2019.101171
43. Silge J, Robinson D. tidytext: Text mining and analysis using tidy data principles in R. *JOSS*. 2016;1(3):37. doi: 10.21105/joss.00037
44. Shcherbakov GD, Bessonov VV. Approaches to the algorithm of analyzing the results of laboratory testing of micro- and macronutrient content of bakery products: Part 2. *Zdorov'e Naseleniya i Sreda Obitaniya*. 2023;31(1):53-59. (In Russ.) doi: 10.35627/2219-5238/2023-31-1-53-59
45. Shcherbakov GD. [Application of the Monte Carlo method for modeling the intake of various nutrients on the example of semi-finished meat products.] In: *Fundamentals of a Healthy Diet and Ways to Prevent Diet-Related Diseases: Proceedings of the Fifth School of Young Scientists with International Participation*,

REFERENCES

1. Drewnowski A, Rehm CD. Consumption of added sugars among US children and adults by food purchase location and food source. *Am J Clin Nutr*. 2014;100(3):901-907. doi: 10.3945/ajcn.114.089458
2. Thompson FE, Subar AF, Loria CM, Reedy JL, Baranowski T. Need for technological innovation in dietary assessment. *J Am Diet Assoc*. 2010;110(1):48-51. doi: 10.1016/j.jada.2009.10.008
3. Costa TPD, Gillespie J, Pelc K, et al. Life cycle assessment tool for food supply chain environmental evaluation. *Sustainability*. 2023;15(1):718. doi: 10.3390/su15010718
4. Delgado A, Issaoui M, Vieira MC, Saraiva de Carvalho I, Fardet A. Food composition databases: Does it matter to human health? *Nutrients*. 2021;13(8):2816. doi: 10.3390/nu13082816
5. Elmadfa I, Meyer AL. Importance of food composition data to nutrition and public health. *Eur J Clin Nutr*. 2010;64(Suppl 3):S4-S7. doi: 10.1038/ejcn.2010.202

- Moscow, November 9–11, 2022. Moscow: Federal Research Center for Nutrition, Biotechnology and Food Safety; 2022:121–122. (In Russ.)
23. Shcherbakov GD, Bessonov VV. Algorithm for analyzing the results of laboratory testing of micro- and macronutrient composition of milk. *Zdorov'e Naseleniya i Sreda Obitaniya*. 2022;30(8):64–72. (In Russ.) doi: 10.35627/2219-5238/2022-30-8-64-72
 24. Akbay A, Elhan A, Ozcan C, Demirtaş S. Hierarchical cluster analysis as an approach for systematic grouping of diet constituents on basis of fatty acid, energy and cholesterol content: Application on consumable lamb products. *Med Hypotheses*. 2000;55(2):147–154. doi: 10.1054/mehy.1999.1038
 25. Ortenzi F, Kolby M, Lawrence M, et al. Limitations of the food compass nutrient profiling system. *J Nutr*. 2023;153(3):610–614. doi: 10.1016/j.tjnut.2023.01.027
 26. Kicherer H, Dittrich M, Grebe L, Scheible C, Klinger R. What you use, not what you do: Automatic classification and similarity detection of recipes. *DKE*. 2018;117:252–263. doi: 10.1016/j.datak.2018.04.004
 27. Subhi MA, Ali SH, Mohammed MA. Vision-based approaches for automatic food recognition and dietary assessment: A survey. *IEEE Access*. 2019;7:35370–35381. doi: 10.1109/ACCESS.2019.2904519
 28. He H, Kong F, Tan J. DietCam: Multiview food recognition using a multikernel SVM. *IEEE J Biomed Health Inform*. 2016;20(3):848–855. doi: 10.1109/JBHI.2015.2419251

Сведения об авторах:

✉ **Щербаков** Григорий Дмитриевич – ведущий инженер лаборатории демографии и эпидемиологии питания; e-mail: sherbakovgrigory@gmail.com; ORCID: <https://orcid.org/0000-0002-9046-6837>.

Бессонов Владимир Владимирович – д.б.н., заведующий лабораторией химии пищевых продуктов; e-mail: bessonov@ion.ru; ORCID: <https://orcid.org/0000-0002-3587-5347>.

Шахвалиева Элина Саид-Аминовна – лаборант-исследователь лаборатории демографии и эпидемиологии питания; e-mail: shelina9558@gmail.com.

Информация о вкладе авторов: концепция и дизайн исследования: Бессонов В.В., Щербаков В.В.; сбор данных: Щербаков В.В.; обработка материала и анализ полученных данных: Щербаков В.В., Шахвалиева Э.С.-А.; обзор литературы, подготовка рукописи: Щербаков В.В., Шахвалиева Э.С.-А. Все авторы ознакомились с результатами работы и одобрили окончательный вариант рукописи.

Соблюдение этических стандартов: исследование не требует представления заключения комитета по биомедицинской этике или иных документов.

Финансирование: исследование выполнено в рамках реализации соглашения с РФФ 23-16-00242 «Разработка программно-аналитического комплекса с web-интерфейсом для обработки, хранения, анализа и верификации данных химического состава пищевой продукции».

Конфликт интересов: авторы декларируют отсутствие явных и потенциальных конфликтов интересов в связи с публикацией данной статьи.

Статья получена: 15.08.23 / Принята к публикации: 11.12.23 / Опубликовано: 29.12.23

Author information:

✉ Grigory D. **Shcherbakov**, Leading Engineer, Laboratory of Demography and Nutritional Epidemiology; e-mail: sherbakovgrigory@gmail.com; ORCID: <https://orcid.org/0000-0002-9046-6837>

Vladimir V. **Bessonov**, Dr. Sci. (Biol.), Head of the Laboratory of Food Chemistry; e-mail: bessonov@ion.ru; ORCID: <https://orcid.org/0000-0002-3587-5347>.

Elina S.-A. **Shakhvaliyeva**, Research Assistant, Laboratory of Demography and Nutritional Epidemiology; e-mail: shelina9558@gmail.com.

Author contributions: study conception and design: Bessonov V.V., Shcherbakov V.V.; data collection: Shcherbakov V.V.; analysis and interpretation of results, literature review, draft manuscript preparation: Shcherbakov V.V., Shakhvaliyeva E.S.-A. All authors reviewed the results and approved the final version of the manuscript.

Compliance with ethical standards: Not applicable.

Funding: The research was carried out within implementation of the agreement with the Russian Science Foundation No. 23-16-00242 “Development of a software and analytical complex with a web interface for processing, storing, analyzing, and verifying data on the chemical composition of food products”.

Conflict of interest: The authors have no conflicts of interest to declare

Received: August 15, 2023 / Accepted: December 11, 2023 / Published: December 29, 2023