УДК 811.581'37+811.581'35

Михалькова Надежда Васильевна
кандидат филологических наук, доцент,
доцент кафедры теории и практики
китайского языка
Белорусский государственный
университет иностранных языков
г. Минск, Беларусь

Nadezhda Mikhalkova
PhD in Philology, Associate Professor,
Associate Professor of the Department
of Theory and Practice of the Chinese
Language
Belarusian State University
of Foreign Languages
Minsk, Belarus
nadezhdakr@yandex.ru

ИСПОЛЬЗОВАНИЕ МЕТОДОВ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ ПРИ ОТБОРЕ РЕПРЕЗЕНТАТИВНОЙ ЧАСТИ МАССИВА ДАННЫХ В ЛИНГВИСТИЧЕСКИХ ИССЛЕДОВАНИЯХ

В статье предлагается пример авторской методики отбора репрезентативной части большого массива лингвистических данных для проведения семантических исследований. Выявляются этапы анализа, необходимые для проведения выборки материала, определяются математические формулы, на которых может быть основан отбор репрезентативной части материала исследования.

Ключевые слова: математическая статистика; массив данных; репрезентативность; лингвистический анализ; выборка.

USING METHODS OF MATHEMATICAL STATISTICS WHEN SELECTING A REPRESENTATIVE PART OF A DATA ARRAY IN LINGUISTIC RESEARCH

The article offers an example of the author's methodology for selecting a representative part of a large array of linguistic data for conducting semantic research. The stages of analysis necessary for conducting a selection of material are identified, and mathematical formulas are determined on which the selection of a representative part of the research material can be based.

Key words: mathematical statistics; data array; representativeness; linguistic analysis; sample.

Отбор материала лингвистических исследований, зависящий, в первую очередь, от поставленных задач, достаточно часто представляет собой сложный процесс определения той части большого массива языковых данных, которая будет являться наиболее репрезентативной областью единиц. Поскольку обработать огромные языковые ресурсы зачастую не представляется возможным ввиду ограничения человеческих возможностей отдельно взятого автора или коллектива авторов, встает вопрос о выборке определенной сферы языковых знаков, которая, несмотря на неполный объем от существующего в языке, будет содержать закономерности, экстраполируемые на весь массив данных.

Таким большим массивом данных для нашего исследования послужил список всех иероглифических знаков китайской письменности, общее число которых насчитывает 79728 единиц [1]. Понимая, что анализ всех представленных единиц является нецелесообразным ввиду уже выявленных на тот период работы в ходе пилотного исследования повторяющихся на каждом новом отрезке материала закономерностей, стала очевидной необходимость разработки методики, которая бы позволила отобрать репрезентативную часть всех иероглифических знаков китайского языка. В то же время, поскольку исследование базируется на выявлении семантических закономерностей, несомненно значимым стало соединение нескольких показателей, как количественных, так и качественных методики отбора материала исследования.

Решение поставленных в исследовании задач выявления семантических связей детерминативов иероглифических знаков китайского языка, а также определения на этом основании принципов их выбора для рассматриваемых сложносоставных единиц китайской письменности предполагало проведение двунаправленного анализа, предусматривающего,

1) с одной стороны, проспективный подход, т.е. «движение от знака к смыслу», соответственно, от семантики детерминатива к семантике производных от него иероглифов, входящих в разные группы в зависимости от детерминатива,

2) с другой стороны, ретроспективный подход, т.е. «от смысла к знаку», соответственно, от семантической области, объединяющей сложносоставные иероглифы китайского языка, к анализу детерминативов, которые входят в состав данных иероглифов.

С целью отбора материала исследования нами была разработана собственная методика, основанная на методах математической статистики и, соответственно, выделении трех критериев ограничения представленного в лексикографических источниках огромного корпуса единиц: *семантическая диверсификация; продуктивность и репрезентативность*. Предложенная методика представлена и осуществлена впервые и может быть применена в будущем для иных лингвистических исследований, особенно в области корпусного анализа больших массивов данных и искусственного интеллекта.

Наличие критерия *семантической диверсификации* вызвано необходимостью анализа различных семантических типов детерминативов сложных иероглифических знаков китайского языка с целью избежания семантической одноаспектности исследования, поскольку исключительно количественный путь (например, частотность) не позволил бы учесть разные семантические группы детерминативов, где потенциально могли быть зафиксированы иные закономерности. Для выявления семантических типов детерминативов нами был установлен их исходный количественный состав и проведена семантическая классификация, базирующаяся на онтологических свойствах обозначаемых детерминативами понятий, из которой впоследствии извлекались данные для показателя *семантической диверсификации*.

Критерий *продуктивность* предполагал анализ числа вхождений детерминатива в сложные иероглифы китайской письменности, на основании чего нами впервые были выделены высоко-, средне- и малопродуктивные детерминативы. Далее показатели *семантической диверсификации* и *продуктивности* сопоставлялись как относительно всей системы детерминативов, так и отдельных выделенных нами в семантической классификации макро- и микрогрупп через критерий *репрезентативности.*

Репрезентативность представляет собой анализ отношения показателя средней величины ($\bar{X}$) в семантической макрогруппе или микрогруппе детерминативов к общему числу производных сложных иероглифов (одновременное взаимодействие двух критериев: *продуктивности* и *семантической диверсификации*) с целью отбора наиболее репрезентативных семантических групп.

Следовательно, первой задачей явилось определение исходного состава детерминативов китайской письменности, который был впоследствии взят за основу анализа.

Согласно статистической науке, чтобы оценить любое явление, не обязательно исследовать всю генеральную совокупность. Возможность экстраполировать выводы о части на целую единицу доказывается математикой. В частности, П. Л. Чебышевым сформулирован «Закон больших чисел» [2], который гласит, что количественные закономерности массовых явлений проявляются только при достаточном числе наблюдений. Следовательно, с одной стороны, чем больше выборка, тем яснее проявляется общая тенденция и компенсируются случайные отклонения.

Вместе с тем, вопрос о том, как определить предел выборки, насколько количественно большой, соответственно, репрезентативной она будет являться и каким образом эту часть исчислить может быть решен с помощью разработанных математических теорем, а также статистических уравнений, которые стали фундаментом для создания формул расчета вероятности ошибки и размера выборки материала исследования для достижения заданной точности.

В основу расчета выборки детерминативов иероглифических знаков китайского языка нами положены методы математической статистики, с помощью которых осуществляется сбор и обработка статистических данных для получения научных и практических выводов. На практике сплошное исследование (каждого объекта из интересующей нас совокупности) проводят крайне редко. К тому же, если эта совокупность содержит большое число объектов или исследование объекта требует нарушения его функционального стандарта, то сплошное исследование нереально. В таких случаях из всей совокупности случайно отбирают ограниченное число объектов и подвергают их исследованию.

При проведении исследования получаются определенные данные, после применения к которым шкалы измерений они становятся числовыми переменными (зависимыми и независимыми). Совокупность элементов, на кото-

рой проводится исследование и которая характеризует все множество (генеральную совокупность), является выборкой. При этом выборка подбирается так, чтобы она представляла все существенные признаки генеральной совокупности, то есть была репрезентативной.

Различают два вида способов отбора: без расчленения генеральной совокупности на части и с расчленением. К первому виду относятся простые случайные отборы (повторные либо бесповторные), когда объекты извлекают по одному из генеральной совокупности. Второй способ отбора включает в себя следующие разновидности соответственно способам расчленения генеральной совокупности. Отбор, при котором объекты отбираются из каждой «типической» части генеральной совокупности, называется типическим.

Если генеральную совокупность делят на число групп, равное объему выборки, с последующим отбором из каждой группы по одному объекту, то такой отбор называется механическим. На практике часто употребляется комбинирование перечисленных способов отбора. Конкретная комбинация способов отбора объектов из генеральной совокупности определяется требованием репрезентативности выборки.

Для того, чтобы наша выборка была репрезентативной, построим выборочное (упорядоченное) распределение величин – детерминативов иероглифических знаков китайской письменности (n=264[1]) от наибольшей к наименьшей, сопровождая данными об их продуктивности.

Рассчитаем выборочное среднее по формуле:

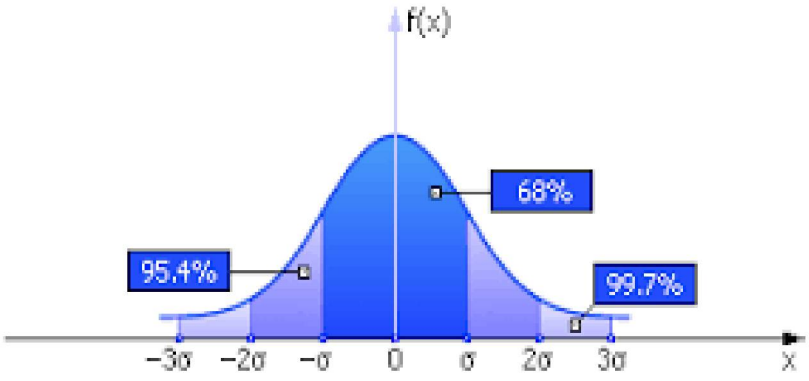
$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k$, где : $\bar{X}$ – выборочная средняя величина; n – количество величин (детерминативы иероглифических знаков китайской письменности); xk – частотность значения показателей. Выражение $\sum_{k=1}^{n} x_k$ соответственно, означает сумму всех X с индексом k от 1 до n и в нашем случае равно 79728 [1].

Таким образом, $\bar{X}$ = 79728/264= 302. Для того, чтобы определить меру того, насколько разбросан набор данных, вычислим стандартное отклонение. Стандартное отклонение соответствует квадратному корню из дисперсии и, наряду с дисперсией, является одной из наиболее часто используемых мер вариабельности признака: $\sigma = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$

[1] Общее число 266 было сокращено до 264, так как для двух детерминативов в используемом иероглифическом словаре полностью отсутствовали данные относительно сложных иероглифов, куда они входят.

Деление суммы квадратов на число степеней свободы n – 1 позволяет сравнивать между собой совокупности, различные по объему. Считается, что дисперсия – более мощный статистический критерий, нежели среднее отклонение, так как больший вклад в дисперсию дают те значения признака, которые расположены дальше от среднего (вклад каждого значения в дисперсию возрастает пропорционально квадрату отклонения от среднего). Стандартное отклонение рассчитаем с помощью приложения Excel. Стандартное отклонение (σ)=499, т.е. разброс наших величин настолько велик, что невозможно определить доверительный интервал, который показывает, в каком диапазоне располагаются результаты выборочных наблюдений.

Согласно правилу трех сигм, данную генеральную совокупность необходимо было бы разделить, удалив часть детерминативов с максимальными и минимальными показателями (рисунок), приблизив выборочный интервал к среднему показателю. Это привело бы к значительным погрешностям выборки ввиду высокого стандартного отклонения. Следовательно, на данном этапе общее число детерминативов невозможно уменьшить, используя только количественный параметр продуктивности.



Нормальное распределение случайной величины

Таким образом, отбор иероглифических данных, согласно критерию – *продуктивности,* показал: наряду с количественной составляющей, необходимо вводить также дополнительный показатель – качественную сторону анализа детерминативов.

Критерий *семантической диверсификации* позволил учесть подобные особенности и максимально охватить разные семантические группы единиц с целью определения как общих для всей системы китайской письменности структурно-семантических закономерностей, так и отличительных свойств отдельных семантических групп детерминативов.

Специфика семантической организация системы детерминативов китайской письменности, а также многоаспектность и разнородность этой системы

единиц подтвердили высказанное ранее предположение о том, что репрезентативная выборка детерминативов должна быть основана не только на количественном, но и на качественном – семантическом параметре, т.е. включать взаимодействие двух критериев: *продуктивности* и *семантической диверсификации*.

Анализ семантических групп детерминативов китайской иероглифической письменности, их продуктивности и деривационного потенциала показал, что существующая широкая диверсификация смысловых компонентов с большой вероятностью может обусловливать наличие различных (в зависимости от семантики детерминатива) закономерностей выбора детерминатива для включения в сложносоставные иероглифы.

Следовательно, отбор детерминативов должен быть осуществлен в соответствии с семантическим критерием, который позволит максимально исключить девиации искомых тенденций.

В каждой из выявленных семантических групп и подгрупп детерминативов были выделены отдельные единицы согласно следующим критериям и процедурам.

Поскольку первый этап расчета выборки позволил определить выборочную среднюю величину по генеральной совокупности, из общей системы нами были исключены детерминативы, показатель *продуктивности* которых не превышал 302 единицы.

Это позволило на втором этапе подсчета ранжировать семантические группы и подгруппы относительно данной величины и не рассматривать семантические подгруппы детерминативов, показатель выборочной средней величины[1] которых не достигал установленного методами математической статистики числа.

Соответственно, в наиболее репрезентативную выборку вошли следующие 12 семантических подгрупп детерминативов: знаки жилищ и их частей ($\bar{X}$=389); знаки природных объектов ($\bar{X}$=428); соматизмы ($\bar{X}$=373); знаки лиц ($\bar{X}$=407); фитонимы ($\bar{X}$=529); зоонимы ($\bar{X}$=368); знаки состояний ($\bar{X}$=909); знаки характеристик ($\bar{X}$=322); знаки видов жидкостей и их состояний ($\bar{X}$ =599); знаки огня ($\bar{X}$=842); знаки твердых веществ ($\bar{X}$=519); знаки сыпучих веществ ($\bar{X}$=433).

Отобранные нами на этом этапе семантические подгруппы содержат 103 детерминатива, что составляет 39 % от всего числа детерминативов в китайской письменной системе. При этом общее число производных сложносоставных иероглифов китайской письменности от данных детерминативов – 55560 единиц, репрезентирующих 69,45 % от всего количества иероглифов в китайском языке.

[1] Выборочная средняя величина подсчитывалась по каждой семантической подгруппе путем поиска среднего арифметического на основании показателей продуктивности детерминативов, входящих в данную подгруппу.

Репрезентативность рассчитывалась исходя из того, что средняя величина подгруппы не должна превышать среднюю величину всех подгрупп – (311,2+880,1)/2=595,65, при этом, например, семантическая подгруппа «Номинации огня» в расчет не вошла, поскольку состоит из 1 детерминатива, учитывая графическую вариативность.По степени репрезентативности могут быть выделены семантические подгруппы, в которых:

1) совпадают или близки по количественному значению показатели средней величины ($\bar{X}$) и общего числа производных сложных иероглифов (например, семантические подгруппы «Соматизмы», «Фитонимы», «Знаки природных объектов» и др.),

2) существуют значительные расхождения показателей средней величины ($\bar{X}$) и общего числа производных сложных иероглифов от каждого детерминатива в данной семантической подгруппе (например, семантическая подгруппа «Знаки сыпучих веществ»).

В отдельных семантических группах имели место детерминативы с показателями максимальной и минимальной продуктивности, например, «Знаки сыпучих веществ»[1], что привело к необходимости проведения дополнительной итоговой процедуры отбора материала исследования с целью исключения детерминативов с минимальной продуктивностью в каждой семантической подгруппе.

Таким образом, в результате применения первых двух критериев из всей системы детерминативов нами были отобраны 12 семантических подгрупп. Поскольку степень репрезентативности детерминативов данных подгрупп является различной, нами был введен показатель выборочного среднего относительно каждой подгруппы, который позволил исключить наименее репрезентативные детерминативы в каждой из подгрупп.

Следовательно, итоговый расчет выборки детерминативов включает следующие единицы (таблица).

Итоговый состав детерминативов
для осуществления анализа сложных иероглифических знаков,
в которые они входят в качестве компонентов

Название группы	Название подгруппы	Детерминативы	Количество	ПД
Объекты неживой природы	Жилища и их части	宀 ‘крыша’ 广 ‘убежище’ 穴 ‘пещера’ 門/门 ‘ворота’	4(5)[2]	689 564 483 616

[1] Детерминатив 卤 ‘соль’ имеет показатель частотности – 1, 土 ‘земля’ – 1592.

[2] В скобках нами указывается число, учитывающее графическую вариативность.

	Природные объекты	日 ‘солнце’ 山 ‘гора’ 月 ‘луна’ 貝/贝 ‘раковина’ 阝 ‘холм’	5(6)	1059 1312 1433 634 1268
	Соматизмы	目 ‘глаз’ 扌/手 ‘рука’ 口 ‘рот’ 足 ‘нога’ 心/忄 ‘сердце’	5(7)	1073 2280 2854 1050 1986
Объекты живой природы	Лица	人/亻 ‘человек’ 女 ‘женщина’ 王 ‘государь’	3(4)	1781 1540 1111
	Фитонимы	木 ‘дерево’ 竹/⺮ ‘бамбук’ 艹 ‘трава’	3(4)	2858 1606 3316
	Зоонимы	虫 ‘червь’ 牛 /牜 ‘бык’ 隹 ‘короткохвостаяптица’	3(4)	1631 426 332
Обозначения абстрактных сущностей	Состояния	疒‘болезнь’	1	909
	Характеристики	饣/食 ‘пища’	1(2)	692
Обозначения веществ	Жидкости	氵/水‘вода’	1(2)	2942
	Огонь	火/灬‘огонь’	1(2)	1685
	Твердые вещества	石‘камень’	1	1024
	Сыпучие вещества	土‘земля’	1	1592
ИТОГО			29(39)	40746

Таким образом, из всей системы детерминативов китайского языка общее число отобранных для анализа детерминативов составило 39 единиц (включая графическую вариативность), образующих 40746 из 79728 иероглифических знаков китайского языка (51,1 %), что составило репрезентативную часть выборки материала исследования.

ЛИТЕРАТУРА

1. 汉字字典 (Ханьцзыцзыдянь). – URL: https://www.zdic.net/. (Дата обращения: 10.03.2022).

2. Нифонтов Н. С. Закон больших чисел и теорема Чебышева [Электронный ресурс] / Н. С. Нифонтов, Е. Ю. Маслова // Вестник Академии знаний. – 2017. – № 20 (1). – Режим доступа: https://cyberleninka.ru/article/n/zakon–bolshih–chisel–i–teorema–chebysheva/viewer. – Дата доступа: 01.04.2023.