О.В. Сенько, Ю.Е. Березкин, С.А Боринская,
А.В. Козьмин, А.В. Кузнецова

ИССЛЕДОВАНИЯ ФОЛЬКЛОРНО-МИФОЛОГИЧЕСКИХ ТРАДИЦИЙ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ[1]

В сравнительных исследованиях по мифологии и фольклору преобладает описание повествовательных традиций в соответствии с лингвистической классификацией. В меньшей степени представлены ареальные исследования — чаще локальные, реже по крупным регионам. Как показывает анализ эмпирического материала, ареалы распространения значительной части фольклорно-мифологических мотивов пересекают языковые границы, а часть мотивов (в том числе сюжетообразующих) встречается на разных континентах. Систематическое исследование сходства фольклорно-мифологических традиций в зависимости от лингвистической или географической близости между соответствующими этническими группами до сих пор не проводилось.

Для того чтобы определить, является языковое родство главным или по крайней мере значимым фактором, определяющим степень

[1] Работа поддержана грантами РФФИ 07–06–00441–а и 08–07–00437–а и программой Президиума РАН «Историко-культурное наследие и духовные ценности России», проект «Древнейшее население Сибири и миграции человека в Новый Свет».

близости традиций друг к другу, авторами была предпринята в своем роде уникальная попытка использовать знания профессионалов-математиков (О.В. Сенько и А.В. Кузнецовой) для обработки огромного массива данных из области гуманитарных наук, а именно данных о встречаемости мифологических мотивов в отдельных традициях. Идея исследования предложена С.А. Боринской. Для исследования была использована база данных «Тематическая классификация и распределение фольклорно-мифологических мотивов по ареалам», созданная Ю.Е. Березкиным на основе анализа почти 6000 публикаций на германских, романских, славянских и прибалтийско-финских языках и некоторых неопубликованных материалов [Березкин 2007; 2009]. На май 2010 г. в ней содержались сведения об ареальном распространении 1570 мотивов, относящихся к 675 традициям. Под мотивом понимаются повторяющиеся образы, эпизоды или их сочетания максимальной протяженности, встречающиеся в двух и более (практически во многих) текстах. В базу данных включались только такие мотивы, которые обнаружены не менее чем в четырех традициях. Под традицией понимается совокупность текстов, записанных у одной или нескольких соседних культурно близких этнических групп. Пополняемый каталог рефератов текстов с указанием традиций, в которых они встречаются, размещен А.В. Козьминым на сайте http://www.ruthenia.ru/folklore/berezkin. Карты распространения мотивов представлены на сайте http://starling.rinet.ru/kozmin/tales/index.php?index=berezkin.

По мере увеличения массива данных происходит дробление ареалов, соответствующих объединению нескольких традиций, с тем чтобы каждая соответствовала лишь одной этноязыковой группе. Пока регионы значительно различаются в отношении полноты собранных данных. Наиболее дробно представлены американские материалы, наименее — африканские. Это объясняется не только неодинаковой изученностью регионов, но и степенью их культурного разнообразия. Материалы по фольклору и мифологии Африки южнее Сахары достаточно однообразны, а по фольклору и мифологии аборигенов Америки различаются очень сильно. Именно поэтому материалы по американским индейцам и эскимосам и были выбраны в качестве своего рода полигона для опробования на них методов математического анализа.

Для статистической обработки материалов создано формализованное описание представленности выявленных мотивов в исследованных традициях. Это таблица, в которой для всех традиций в би-

нарной форме фиксируется наличие или отсутствие каждого мотива в проанализированных источниках.

Ставились следующие цели анализа:

— оценка взаимной близости традиций по всей совокупности мотивов;

— выявление групп традиций, однородных по набору встречающихся мотивов;

— оценка статистической достоверности различий между заранее заданными группами традиций (например, соответствующих одной или нескольким языковым семьям) по полной совокупности представленных в них мотивов;

— выявление отдельных мотивов или их сочетаний, демонстрирующих статистически достоверные различия по встречаемости в заранее заданных группах традиций.

МЕТОДЫ АНАЛИЗА

Попарная близость между традициями. Любым двум традициям *T[i]* и *T[j]* может быть сопоставлена функция расстояния S(*T[i], T[j]*), отражающая близость их мотивов. Формально произвольной традиции *T[j]* может быть сопоставлена случайная индикаторная функция *t[j]*, заданная на множестве мотивов и принимающая значения 0 и 1 в зависимости от того, зарегистрировано или нет наличие мотива. Следует подчеркнуть, что наличие 0 в некоторой позиции не обязательно достоверно свидетельствует о реальном отсутствии в соответствующей традиции соответствующего мотива, поскольку некоторые традиции изучены неполно. Последнее обстоятельство не позволяло использовать в качестве функции близости стандартные метрики Евклида или Хэмминга, которые предполагают суммирование совпадений по всем мотивам, поскольку это приводит к установлению высокой близости между любыми двумя слабо исследованными традициями. В связи с этим были выдвинуты альтернативные функции расстояния.

Предположим, что общее количество мотивов в исследуемой базе составляет *N*.

1) Функция *Sk*(*T[i], T[j])=1-0.5*{K*(*t[i],t[j])+1}* представляет собой значение обычного (пирсоновского) коэффициента корреляции, формально примененного для сравнения бинарных функций *t[i]* и *t[j].*

2) Функция $Sc(T[i], T[j])=1-0.5*\{ k*C(t[i],t[j]+1\}/N$, где $C(t[i],t[j])$ представляет собой величину статистики критерия Хи-квадрат при проверке равенства распределения значений функции $t[i]$ в группах, формируемых индикаторной функцией $t[j]$. Коэффициент k принимает значение 1, если доля мотивов с $t[i]=1$ при $t[j]=1$ превышает долю мотивов с $t[i]=1$ при $t[j]=0$, и k принимает значение -1 в противном случае.

Обе функция расстояния становятся равными 0 при полной тождественности описаний двух традиций, оказываются равными 0.5 при условии статистической независимости встречаемости мотивов в двух традициях и оказываются равными 1 при полном несовпадении встречаемости мотивов.

Данные функции близости позволяют более адекватно описывать сходство между традициями при слабой степени их заполненности. Вместе с тем влияние количества представленных мотивов в сравниваемых традициях на степень сходства сохраняется также и для них. Для иллюстрации отметим, что, случайным образом и независимо, исключая мотивы в двух очень близких традициях, мы можем придти к ситуации, когда подмножества оставшихся нескольких мотивов имеют небольшое пересечение или не пересекаются вообще.

Для того чтобы оценить влияние степени наполнения традиций мотивами (далее — наполненность) на величины функций взаимной близости между ними было проведено следующее исследование.

Пусть $m[i]$ — число мотивов, зарегистрированных в традиции $T[i]$. Случайным образом из 56 616 всевозможных пар, содержащихся в базе данных традиций, для дальнейшего анализа было отобрано 10 %. Для этих отобранных пар с помощью методов корреляционного анализа исследовалась зависимость $Sk(T[i],T[j])$ от $m[i],m[j]$ и квадратного корня из их произведения $sqrt(m[i]*m[j])$.

Величины коэффициентов корреляции, характеризующие линейную зависимость

$1-Sk(T[i],T[j])= 0.5*\{K(T[i],T[j])+1\}$ от параметров: $mleft$ — наполненность традиции в левой позиции выбранной пары; $mright$ — наполненность традиции в правой позиции выбранной пары; $sqrt(mleft*mright)$ и числа $mrand$, случайным образом и равновероятно выбранного из пары $(mleft*mright)$, показаны в табл. 1. Видно, что относительно слабая, но статистически значимая линейная связь расстояния между традициями и степенью их изученности действительно существует.

Таблица 1

Корреляция расстояния между традициями и наполненностями

mleft	*mright*	*mrand*	*sqrt(mleft*mright)*
0.17	0.15	0.16	0.26

При этом оказалось, что выраженность зависимости 1-*Sk*(*T[i],T[j])* от степени изученности традиций снижается по мере увеличения последней. В строках табл. 2 представлены коэффициенты корреляции, рассчитанные по парам традиций, представленных количеством мотивов в каждой не ниже 50, 70 и 100.

Таблица 2

Корреляция расстояния между двумя традициями и их наполненностью для разных интервалов последней

	mleft	*mright*	*mrand*	*sqrt(mleft*mright)*
≥50	0.11	0.09	0.11	0.15
≥70	0.10	0.08	0.10	0.13
≥100	0.06	0.06	0.05	0.08

Зависимость 1-*Sk*(*T[i],T[j])* от наполненности m_{rand} наглядно показана на рис. 1.

Тангенс угла наклона прямой, описывающей линейную зависимость 1-*Sk*(*T[i],T[j])* от *mrand*, составляет $7{,}2 \cdot 10^{-5}$ на мотив (менее 0.01 на 100 мотивов). Иными словами, при сравнении близости традиций *T[1]* и *T[2]* к традиции *T* эффект влияния степени наполненности позволяет уверенно объяснить $|Sk(T[1],T)-Sk(T[2],T)|<0.01$ при $m_1 - m_2 = 100$. Стандартное отклонение *Sk*(*T[i],T[j])* на множестве всевозможных пар традиций составляет около 0.0495. Отсюда может быть сделан вывод, что функция близости *Sk*(*T[i],T[j])* может являться эффективным инструментом для выявления различий между традициями именно по составу мотивов.

Выявление однородных групп традиций. Для выявления групп традиций с близким характером встречаемости мифологических мотивов применен статистический метод иерархической группировки. При анализе выбранного массива данных (в рассматриваемом случае — все внесенные в базу традиции Нового Света) на каждом шаге происходит объединение кластеров с максимальным значением

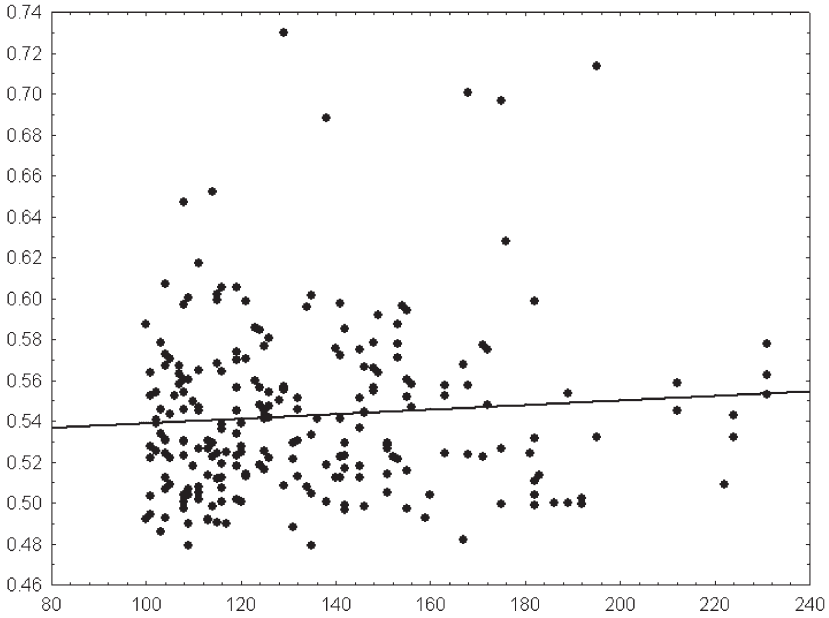

Рис. 1. Зависимость 1-*Sk*(***T[i], T[j]***) от наполненности *mrand*

усредненной (по всем парам объектов из разных кластеров) функции близости. Число формируемых кластеров зависит от заданной степени близости. Полученные результаты показывают, что в кластерах, образованных по сходству мифологических мотивов, представлены, как правило, традиции географически близкие. Территориальное распределение 12 кластеров для американского континента представлено на рис. 2. Видно, что за единственным исключением (гуахиро северо-западной Венесуэле оказались вместе с североамериканскими индейцами) кластеры представляют собой географически компактные группы.

Изменения заданной степени близости объединяемых в кластер традиций приводят к сокращению или увеличению числа ареальных кластеров. Так, для одной лишь Северной Америки число кластеров можно сократить до четырех (север и северо-запад, запад, восток, юг). Однако во всех случаях корреляция с языковыми семьями отсутствует. Например, в северо-западный кластер входят эскоалеуты, на-дене и вакаши, языки которых родственны друг другу не более, чем любые взятые наугад языки Евразии и Нового Света.

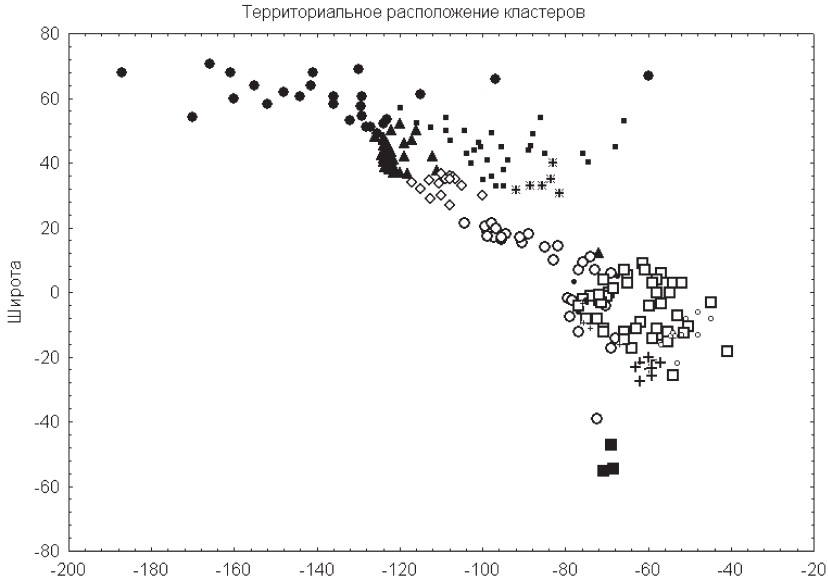

Рис 2. Территориальное распределение кластеров близких традиций для американского континента. Обозначения:

- ● Американская Арктика и северо-запад Северной Америки: алеуты и эскимосы, на-дене, хайда, вакаши, а также цимшиан (макросемья пенути) и беллакула (сэлишская семья).
- ■ Северо-восток, Средний Запад, Равнины: сиу, алгонкины, кэддо, северные ирокезы.
- ✳ Юго-восток: мускоги, ючи, алгонкины, южные ирокезы, языковые изоляты низовьев Миссисипи.
- ▲ Плато — Калифорния: сэлиши, пенути, хока, северные юто-ацтеки, тихоокеанские атапаски, а также ряд языковых изолятов и микросемей.
- ◇ Юго-запад: хока, северные и южные юто-ацтеки, керес, южные атапаски, таньо, зуньи.
- ○ Мезоамерика — Анды: южные юто-ацтеки, майя, ото-манге, тотонаки, михе-соке, чибча, кечуа, такана, различные микро-семьи и изоляты.
- □ Амазония: араваки, карибы, макро-тупи, тукано, яномамо, пано, различные микросемьи и изоляты.
- ○ Бразильское нагорье: макро-же, тупи, араваки, карибы, ряд изолятов.
- ┼ Чако: гуайкуру, самуко, матако-матагуайо, маскои.
- • Северо-запад Южной Америки: чибча, салива, ваорани, чаяуита.
- + Прендандские районы Амазонии: араваки, ряд изолятов и микросемей.
- ■ Южный Конус: чон, яганы.

Статистическая достоверность различий между группами традиций по совокупности мотивов. Для оценки статистической достоверности различий между двумя непересекающимися группами традиций *G1* и *G2* использовались функционал качества *Fm*(*G1*,*G2)* и перестановочный тест.

Функционал *Fm*(*G1*,*G2)* представляет собой произведение *(m1* m2)*[Dinter- Dall]*, где

mi — число традиций в группе *Gi, i=1,2*;

Dinter — среднее расстояние в смысле *Sk*(*T[i], T[j])* или *Sc*(*T[i], T[j])* между традициями из двух разных групп;

Dall — среднее расстояние между традициями из объединения двух групп.

Для оценки достоверности различий между двумя группами значение функционал *Fm* на реальных данных сравнивается со значением этого же функционала на выборке, полученной из исходной реальной выборки путем случайных перестановок индикаторной функции классов. Доля перестановок, для которых значение *Fm* превышает значение *Fm* на реальных данных, принимается в качестве статистической меры достоверности различий (**p-значения**). Следует отметить высокую степень обоснованности использования перестановочного теста, не требующего предположений о характере вероятностного распределения и не имеющего ограничений по размеру выборки [Журавлев, Рязанов, Сенько 2006; Сенько 2003]. Предложенный подход был использован для сравнения уровня сходства мотивов в традициях коренных обитателей Америки, языки которых относятся к разным семьям. Результаты представлены в таблице 3 (уровень значимости *P* рассчитан с помощью перестановочного теста).

Таблица 3

Различия между двумя лингвистическими группами традиций по совокупности мотивов

Группа 1	Группа 2	Среднее расстояние			*P*
		между традициями внутри группы1	между традициями внутри группы 2	между традициями групп 1 и 2	
Алгонкины	Сиу	0.37	0.35	0.37	0.14
Эскоалеуты	Карибы	0.34	0.39	0.47	0
Юж.юто-ацтеки	Майя	0.38	0.32	0.34	0.51

Кечуа	Макро-тупи	0.39	0.39	0.43	0.002
Майя	Кечуа	0.32	0.39	0.41	0.025

Из таблицы видно, что средние расстояния по сходству совокупности мотивов между географически близкими языковыми группами алгонкинов и сиу, южных юто-ацтеков и майя не превышают таковых расстояний внутри этих групп.

Различия между группами по отдельным мотивам. Для количественной оценки степени различий представленности сюжета в непересекающихся группах традиций *G1* и *G2* использовались функционал качества $Fq(G1,G2)$ и перестановочный тест. Функционал $Fq(G1,G2)$, по сути, представляет собой величину статистики критерия Хи-квадрат при проверке гипотезы об одинаковой встречаемости сюжета в группах *G1* и *G2*. Использование перестановочного теста совершенно аналогично описанному в предыдущем параграфе.

Оценка близости двух групп фольклорных традиций по степени коррелированности расстояний от них. Дополнительным способом оценки близости двух групп фольклорных традиций *G1* и *G2* является вычисление коэффициента линейной корреляции $K(R1,R2)$ между двумя соответствующими им функциями $R1(T)$ и $R2(T)$, заданными на некотором множестве традиций $\{T\}$. При этом $Ri(T)$ представляет собой среднее расстояние между традицией *T* и группой традиций *Gi*. Следует отметить, что использование для оценки степени сходства традиций коэффициентов $K(R1,R2)$ позволяет в значительной степени избежать искажающего влияния степени полноты описания традиций (наполненности, т.е. количества представленных для нее в базе мотивов), а также эффективнее показать статистическую достоверность наличия сходства и наглядно представить результаты исследования.

На рис. 3 представлена зависимость расстояния до совокупности традиций языковой семьи майя от расстояния до фольклорной традиции Китая, полученная на множестве всех американских традиций. Коэффициент корреляции $K(R1,R2)$. В табл. 4 представлены значения аналогичным образом рассчитанных коэффициентов, характеризующих связь майя с некоторыми внеамериканскими традициями.

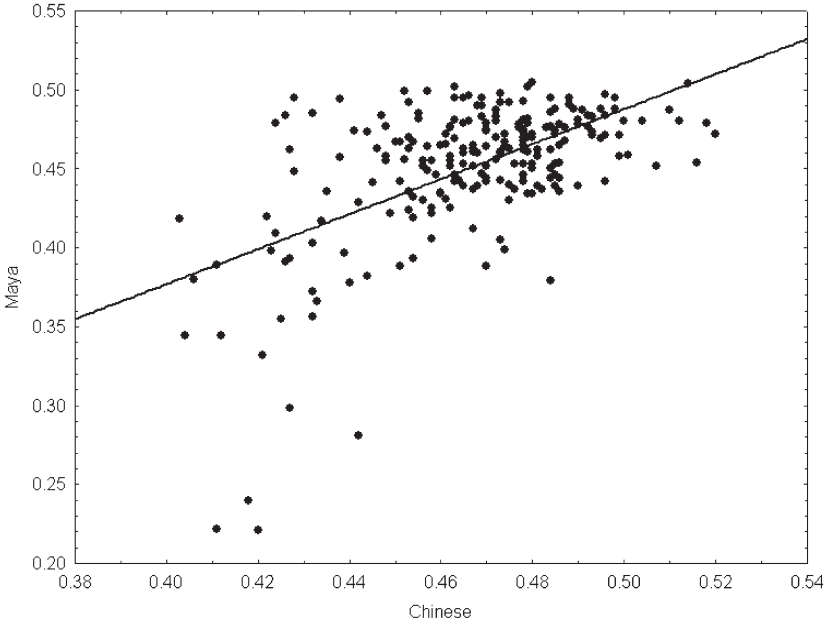

Рис. 3. Для американских традиций представлена зависимость расстояния до традиций языковой семьи майя от расстояния до традиции Китая

Таблица 4

Связь языковой семьи майя с некоторыми внеамериканскими традициями

Традиция	*K(R1,R2)*.
Хадза, Сандаве (Восточная Африка)	0.31
Чукчи	-0.17
Байкало-амурские эвенки	0.01
Айну Сахалина и Хоккайдо	-0.09
Папуасы Новой Гвинеи	0.31
Китай	0.57

ВЫВОДЫ

Проведенное исследование показало, что лингвистическое родство не является определяющим фактором, влияющим на сходство

наборов мотивов, характерных для изученных традиций. Превалирование фактора географической близости над фактором языкового родства не вызывает сомнений. Строго говоря, интуитивно и эмпирически это было ясно с самого начала работы по созданию базы данных фольклора и мифологии. Существенно, однако, что применение достаточно изощренной математической процедуры данное мнение подтверждает. Кроме того, исследование продемонстрировало высокую устойчивость результатов анализа при модификации базы данных.

Сопоставление полученных ареальных кластеров с прочими факторами, которые гипотетически могли бы влиять на распространение определенных мотивов в пределах определенных территорий, не дает сколько-нибудь четкой картины. Группы с присваивающей экономикой в общем и целом оказываются в других кластерах, нежели земледельцы, однако различия между отдельными «охотничье-собирательскими» кластерами или отдельными «земледельческими» кластерами не больше и не меньше, чем между охотниками-собирателями и земледельцами. То же касается корреляции с разного рода природно-климатическими факторами и с различиями в уровне социально-политической организации.

Итак, с одной стороны, проведенное исследование, как уже было сказано, показало, что лингвистическое родство не является определяющим фактором, влияющим на сходство наборов мотивов, характерных для изученных традиций. Для многих языковых семей степень близости входящих в них традиций немного выше, чем усредненная степень близости между семьями, но это вызвано лишь принадлежностью семей к разным ареальным кластерам. Для соседних семей, принадлежащих к одному кластеру (алгонкины и сиу, карибы и араваки) различия нулевые.

С другой стороны, вовсе не исключено, что в основе своей ареальные кластеры соответствуют языковым объединениям, но не тем, которые зафиксированы в Америке в период после начала европейских контактов, а тем, которые участвовали в начальном заселении Нового Света. Существенно, что отдельные американские кластеры сходны с разными азиатскими и океанийскими традициями (от Меланезии до северной Сибири). Это подтверждает предварительный вывод об участии в заселении Нового Света разнокультурных азиатских популяций. Южно-американские кластеры близки Меланезии, а некоторые (особенно Чако и Южный Конус) имеют достаточно высокий уровень близости к Австралии и к Африке южнее Сахары.

Представляется вероятным, что степень устойчивости фольклорно-мифологических традиций, способных неопределенно долго сохранять свойственные им наборы мотивов, значительно выше степени устойчивости языков. Первоначальная языковая карта Нового Света могла совпадать с картой распространения фольклорно-мифологических мотивов. Однако за прошедшие десять и более тысячелетий в ходе сотен и тысяч крупных и мелких миграций эта карта неузнаваемо изменилась, поскольку многие ареальные популяции многократно переходили с одного языка на другой. Конкретные случаи подобного рода хорошо известны для Нового Света, а для Сибири прослежены такие изменения языковой принадлежности населения, которые охватывали огромные территории и происходили на протяжении веков, а не тысячелетий (например, [Долгих 1952]). Учитывая вероятную локализацию прародин отдельных языковых семей Нового Света (алгонкинской, сиу, юто-ацтекской, атапаскской, карибской, аравакской, чибча и других), можно уверенно утверждать, что всего лишь 4–5 тыс. л.н. языковая карта Америки радикально отличалась от существовавшей там к 1500 г. н.э.

Миграции редко (если вообще когда-либо) приводили к полному вытеснению предшествовавшего населения. Скорее это было медленное просачивание нового населения, которое, оказавшись в новой природной среде, заимствовало многие элементы культуры от более ранних обитателей. Сверхдальних, трансконтинентальных миграций в период после начального освоения Нового Света, скорее всего, уже не было. Соответственно, можно предположить, что ареальные фольклорно-мифологические традиции сохранялись, но их носители утрачивали былое языковое единство.

Библиография

Березкин Ю.Е. Мифы заселяют Америку. Ареальное распределение фольклорных мотивов и ранние миграции в Новый Свет. М.: О.Г.И., 2007.

Березкин Ю.Е. Мифы Старого и Нового Света. М.: АСТ: Астрель, 2009.

Долгих Б.О. Происхождение нганасанов // Сибирский этнографический сборник. М.; Л., (Труды института этнографии. Нов. сер.). 1952. Т. 18. С. 5–87.

Сенько О.В. Перестановочный тест в методе оптимальных разбиений // Журнал вычислительной математики и математической физики. 2003. № 9. С. 1438–1447.

Журавлев Ю.И., Рязанов В.В., Сенько О.В. Распознавание. Математические методы. Программная система. Применения. Москва: Фазис, 2006.