digitalorientalia
Цифровое востоковедение
DOI 10.31696/S278240120025095-7
VOL. 3 | No. 1-2 | 2023

Цифровые методы в лингвистике | Digital Methods in Linguistics

Аудиотестирование как способ классификации даргинских идиомов

Audio Testing as a Method of Classification of Dargin Idioms

Магомедов Ахмед Гусенович

НИУ ВШЭ, г. Москва, студент магистратуры «Цифровые методы в гуманитарных науках»

E-mail: akhmed1@yandex.ru
ORCID: 0000-0002-6990-9859

Akhmed G. Magomedov

Student of MP «Digital Humanities» HSE University

Резюме. В статье кратко рассматривается оценка видов аудиотестирования, которые применялись и которые могут применяться для классификации даргинских идиомов, так как исследователями до сих пор точно не принята единая классификация даргинских идиомов (языков и диалектов). Дана критическая оценка методологии, исследованы возможности аудиотестирования методом RTT и с помощью «Фильма о грушах». Также в статье описана методология для создания теоретической базы для тестирования даргинских идиомов, включающая корпусы диалектов с аудиофайлами, созданные Научно-учебной лабораторией по формальным моделям в лингвистике Школы лингвистики «НИУ ВШЭ».

Ключевые слова: тестология, аудиотестирование, языковые тесты, даргинские языки, дагестанские языки, RTT

Abstract. The article briefly discusses the assessment of the types of audio testing that have been used and can be used to classify Dargin idioms, since researchers have not yet accurately adopted a common classification of Dargin idioms (languages and dialects). A critical assessment of the methodology is given, the possibilities of audio testing using the RTT method and the "Pear stories" are investigated. The article also describes a methodology for creating a theoretical basis for testing Dargin idioms, including corpus of dialects with audio files created by the Laboratory for Formal Models in Linguistics of the HSE School of Linguistics.

Keywords: testology, audio testing, language tests, Dargin languages, Dagestanian languages, RTT

О даргинских идиомах

Республика Дагестан славится своим этническим разнообразием. По разным оценкам, в Дагестане насчитывается более 30 народностей. К числу самых многочисленных относятся даргинцы. Однако как такового единого даргинского языка не существует. Говоря о даргинцах, упоминают термин «даргинские языки», так как это семейство языков и диалектов, объединенных условно единой языковой базой.

Споры вокруг классификации даргинских языков не утихают по сей день. Лингвисты-кавказоведы расходятся во мнениях по следующим пунктам:

а) разграничение понятий «язык» и «диалект», так как некоторые идиомы именуют и языком, и диалектом. Например, кадарский и муиринский идиомы;

б) разграничение идиомов по группам.

Опираясь на грамматический принцип, лингвист-даргиновед Муталов Р.О. выделяет северную подгруппу даргинских языков из 10 идиомов и южную подгруппу, состоящую из 16 (17) идиомов [1]. Помимо прочего, также существует отдельный литературный даргинский язык (Dargwa language), на котором публикуется литература и местные периодические издания.

Севернодаргинские языки и их диалекты	**Южнодаргинские языки и их диалекты**
1.1. акушинский (диалекты): акушинский, гашшиминский, губденский, кадарский, мекегинский, мугинский, муиринский, мюрегинский, урахинский 1.2. мегебский	2.1 сирхя-цудахарский: амузги-ширинский, амухский, бутринский, ицаринский, кункинский, санжинский, сирхинский, тантынский, усишинский, цудахарский, худуцский 2.2. кайтагский: верхнекайтагский, нижнекайтагский, чахри-санакаринский, шаринский 2.3. кубачинский: аштынский, кубачинский 2.4. чирагский

Таблица 1. Классификация даргинских идиомов по Муталову Р.О. [1].
Нумерация – Магомедов А.Г.

Тестирование пересказа записанного текста (RTT Retelling)

Помимо исследований грамматических различий, исследователи также опираются на лексико-фонетический принцип. Так, в 2019 году группа исследователей в попытке сравнить взаимопонимание носителей различных даргинских идиомов описала эксперимент [3], основанный на тестировании пересказа записанного аудиотекста по методу Recorded Text Testing Retelling (RTT Retelling)[1]. Данный метод хорошо себя зарекомендовал и заключается в тестировании понимания устного текста носителями разных идиомов, что помогает оценить в процентном соотношении степень понимания носителями одного диалекта носителей другого диалекта. Для этого респондентам (носителям даргинских идиомов) предлагалось прослушать одноминутный аудиорассказ (состоящий из 7 сегментов) на другом даргинском идиоме, пересказать услышанное на своем родном диалекте и перевести сказанное на русский язык (для оценки понимания рассказа). После перевода текст сопоставляется с первоначальным переводом рассказа по проценту совпадений ключевых слов и словосочетаний. Успешность коммуникации оценивается в процентном измерении: 85% и выше – успешная коммуникация, 71 – 85% – на границе взаимопонимания, ниже 70% – сильное различие идиомов, взаимопонимание затруднительное.

По результатам исследования, было выявлено, что, к примеру, акушинцы (1.1) понимают муиринцев (1.1) лучше (62%), чем представителей южнодаргинских языков: цудахарцев (2.1) – 52%, кайтагцев (2.2) – 29%, кубачинцев (2.3) – 18%.

[1] Авторы: Casad [4], Blair [5], Grimes [6]

Однако при этом кубачинцы, будучи в южной подгруппе, лучше поняли речь муиринцев из северной подгруппы (55%), чем кайтагцев (48%) и цудахарцев (50%), что по-своему усложняет классификацию даргинских идиомов, которая может видоизменяться с лексико-фонетической точки зрения. Но для подробного анализа требуется создание корпусов слов даргинских идиомов и их сравнение между собой. Пересказ аудиорассказа длиной в 1 минуту не является достаточно репрезентативным и не может гарантировать высокую валидность тестирования. К тому же, рассказчики были представителями лишь 4 идиомов: акушинский, кайтагский, кубачинский, цудахарский. А респондентами выступили представители 10 идиомов. Однако у половины представленных идиомов было всего по одному респонденту, что снижает объективность исследования (см. Таб. 2).

Диалекты респондентов	**Пол**		**Возраст**		
	Женский	**Мужской**	**<30**	**30–50**	**>50**
Акушинский (1.1)	6	7	7	3	3
Урахинский (1.1)	2	0	1	0	0
Мюрегинский, губденский (1.1)	1	0	1	0	0
Гапшиминский (1.1)	6	3	1	2	5
Муиринский (1.1)	1	0	0	1	0
Цудахарский (2.1)	2	6	2	4	2
Тантынский (2.1)	1	0	0	0	1
Сирхинский (2.1)	3	1	0	2	0
Кайтагский (2.2)	1	0	0	0	1
Кубачинский, аштынский (2.3)	1	0	0	1	0
Итого	**24**	**17**	**12**	**13**	**12**

Таблица 2. Респонденты, участвовавшие в исследовании В. Малышева и др. [3].

В оригинале метод включает в себя 3-минутный аудиорассказ, разделенный на 10-12 сегментов (по 1-2 предложениям) [6].

Подготовка к пересказу включает в себя:

1. Запись на диктофон короткого рассказа носителем диалекта на повседневные (бытовые) темы,

2. Транскрипцию и перевод (буквальный и идиоматический) полученного текста, проверка корректности перевода на язык исследователя (так как исследователи зачастую не являются носителями языка/диалекта),

3. Разделение с помощью коротких пауз на 10-12 контекстных сегментов для более точного анализа пересказа респондентом отдельных предложений,

4. Определение ключевых словосочетаний или слов для каждого сегмента,

5. Валидацию достаточного лексического разнообразия текстов (тексты должны быть в достаточной мере содержательными, с использованием разнообразных слов и словосочетаний).

Результат оценивается следующим образом:

В каждом сегменте выделяется 4 ключевых словосочетания, каждое из которых оценивается в 1 балл. Если респондент при пересказе передает верный смысл словосочета-

ния (допускается использование близких синонимов), то он получает 1 балл. Если частично (например, правильное подлежащее и неправильное сказуемое), то 0,5 балла. Максимум баллов за 1 сегмент – 4 балла. Набранные за каждый из 10-12 сегментов баллы суммируются в итоговый результат.

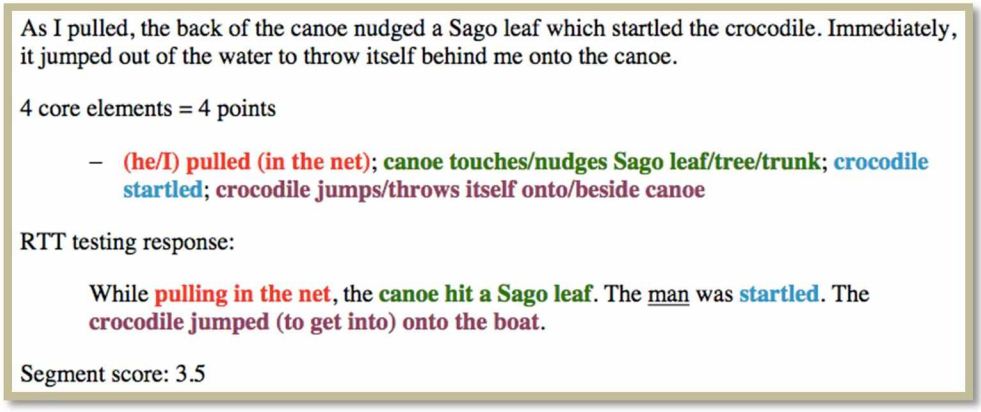

Рис. 1. Тестирование RTT на примере западного диалекта языка Сентани (провинция Папуа, Индонезия) [7]. Первое предложение – это сегмент оригинального высказывания. Далее выделяются 4 ключевых словосочетания, и ниже приводится интерпретация услышанного респондентом, в которой была допущена ошибка в одном из ключевых элементов.

Основным преимуществом Тестирования пересказа записанного текста (RTT Retelling) является тот факт, что в данном тестировании проверяется понимание всего текста, а не только выбранных фрагментов. Второе важное преимущество заключается в том, что для многих людей пересказ услышанной истории является более комфортным, чем ответы на вопросы, которые могут быть расценены респондентом как классическое тестирование и могут влиять на уровень волнения и точность высказываний. Дополнительным преимуществом является то, что этот метод не требует проектирования вопросов понимания и перевода этих вопросов в речевые разновидности исследуемых сообществ.

Однако для более высокой валидности и надежности тестирования RTT рекомендуется предлагать респондентам аудиорассказы от не менее чем 3 носителей идиома. Иными словами, каждому респонденту необходимо прослушать по 3 записи на отдельном идиоме. А рекомендуемое минимальное количество респондентов-носителей каждого отдельно взятого идиома – 5 человек. Данная рекомендация обусловлена тем, что:

1. Словарный запас или используемый лексикон носителя может содержать много слов, которые могут быть знакомы респондентам. Однако ограничение в 3 минуты в недостаточной мере может предложить вокабуляр для оценивания.

2. Использование аудиозаписей минимум 2 человек позволит значительно увеличить тестируемый вокабуляр и повторно проверить слова, понимание которых могло вызвать затруднение у респондентов. Ведь в зависимости от контекста уровень понимания отдельных слов может меняться.

3. Респонденты имеют разный словарный запас. Тем более, что порой некоторые слова у представителей одного и того же идиома могут отличаться друг от друга.

«Фильм о грушах»

Для тестирования различий в диалектах также используют методику, которая была предложена в 1980 году в коллективной монографии под редакцией Уоллеса Чейфа «Рассказы о грушах: когнитивные, культурные и языковые аспекты порождения повествования». В данном исследовании носителям разных языков предлагалось просмотреть шестиминутный видеоролик под названием «Фильм о грушах», который был снят специально для исследования [8]. В фильме фермер собирает с дерева груши, которые у него крадет мальчик на велосипеде. Также в фильме появляются другие обитатели села. Герои фильма не произносят ни единого слова. Помимо этого, в кадре появляется игрушка, у которой нет специального названия и которая представляет собой ракетку для пинг-понга с привязанным мячиком. Испытуемым необходимо было пересказать своими словами увиденное (подобно репортажу). Записи проводились с испытуемыми разных возрастов, а также с различными временными интервалами между просмотром фильма и пересказом. Данная методика послужила появлению «Китайских рассказов о грушах»[2], сайта с пересказами «Фильма о грушах» на семи китайских диалектах, что активно используется учеными в исследованиях анализа дискурса.

«Фильм о грушах» может быть использован в исследованиях по классификации даргинских идиомов, так как:

1. представляет собой единый набор образов для лексического тестирования; имеется возможность сравнить, как именуют явления или действия, как грамматически выстраивают предложения представители разных идиомов;

2. в отличие от тестирования RTT не требует задействования двух групп испытуемых (рассказчиков и респондентов);

3. тестирует лексическое разнообразие, «богатство» словарного запаса и корректность грамматики.

Помимо прочего, «Фильм о грушах» отлично подходит для использования в образовательных учреждениях, где преподают литературный даргинский язык как способ языкового контроля обучающихся.

Возможности для аудиотестирования

В данный момент на базе Научно-учебной лаборатории по формальным моделям в лингвистике Школы лингвистики «НИУ ВШЭ» ведется проект «Вариативность в дискурсе и словаре: исследование близкородственных языков цифровыми методами»[3] (поддержан РНФ). В рамках проекта был создан «Dargwa Dictionary Project», который представляет собой сравнительную базу даргинских языков и диалектов для лингвистов и носителей языка. С помощью поисковой строки выполняется поиск по словарной базе. Слова можно вводить на русском или английском языках. Результат выдается на одном выбранном или на всех пяти диалектах в базе (на данный момент): акушинский (1.1), кадарский (1.1), муиринский (1.1), ицаринский (2.1), тантынский (2.1). Также в ближайшее время база будет дополнена урахинским диалектом (1.1).

У большинства слов при переходе на страницу «К статье слова» имеются также аудиофайл с произношением слова и словарная справка.

Помимо функции словаря и сравнения слов между диалектами, данная база может служить источником для создания тестовой среды. Однако ее следует дополнить словарем литературного даргинского языка. Наиболее эффективным видом тестирования

[2] The Chinese Pear Stories. Режим доступа [URL]: www.pearstories.org

[3] Исследование даргинских языков. Режим доступа [URL]: https://hum.hse.ru/fml/dargwa

представляется тест на знание литературного даргинского с 4 вариантами ответов (подбор верного эквивалента). В качестве вопросов и вариантов ответов необходимо использовать 3 варианта теста: с русского на литературный даргинский и наоборот, с русского на даргинский идиом и наоборот, с литературного даргинского на идиом[4] тестируемой группы и наоборот. Такой способ позволит выявить, при каких условиях даргинцы понимают те или иные слова на литературном языке и на родном диалекте/языке. Альтернативно данные тесты можно создать на базе аудиофайлов с произнесенными словами, так как рамках вышеупомянутого проекта слова снабжаются аудиоматериалами. Проект открыт для носителей даргинских идиомов.

Рис. 2. Интерфейс сайта проекта «Dargwa Dictionary Project».
Для примера был выполнен поиск по глаголу «вставать»[5].

[4] При тестировании акушинского диалекта стоит также учитывать тот факт, что он является основой литературного даргинского.

[5] Dargwa Dictionary Project. Режим доступа [URL]: http://lingconlab.ru/dargwa_dict/

Заключение

На данный момент возможности применения аудиотестирования в даргинских идиомах недостаточно изучены, требуют совершенствования использования перечисленных методик, участия широкого круга испытуемых и привлечения источников финансирования. В первую очередь, необходимо собрать испытуемую группу в достаточном количестве для записи аудиоматериалов по методу RTT Retelling.

Целесообразной для создания полноценных текстовых тестов для языкового контроля представляется дальнейшая разработка корпуса в виде рабочей базы литературного даргинского и даргинских идиомов. На данный момент единственным академическим словарем по даргинским языкам является Даргинско-русский словарь (литературный даргинский)[6], основа для которого была заложена дагестанским лингвистом Абдуллаевым З.Г.

Для более эффективной работы требуется трансформация данного словаря из простого машиночитаемого вида (DOC, PDF) в структурированный набор данных (например, TEI[7]). Автором была предпринята попытка создания инструмента для трансформации словарей в набор данных в рамках магистерской диссертации[8]. На примере выборки из 100 слов Даргинско-русского словаря была разработана методика цифровизации словарей на основе алгоритма Earley. Данная методика позволяет трансформировать не только весь Даргинско-русский словарь в структурированный набор данных (с дальнейшей конвертацией в другие форматы), но и словари на других языках, в том числе с редкими алфавитами.

Литература

1. Муталов Р.О. Глагол даргинского языка. Махачкала: Изд.-полиграф. центр ДГУ. 2002. 216 с.
2. Муталов Р.О. Классификация даргинских языков и диалектов [Электронный ресурс] // Социолингвистика. 2021. № 3 (7). С. 8–25.
3. Malyshev, V., Malysheva, V., Gutz, A., Novaya, I., Panina, A., Yurkova, A., Clifton, J.M., Tiessen, C. (2019). The Sociolinguistic Situation of the Dargwa in Dagestan. SIL Electronic Survey Report 2019-011. Dallas: SIL International. 51 p.
4. Casad, Eugene. 1974. Dialect intelligibility testing. Norman, OK: Summer Institute of Linguistics.
5. Blair, Frank. 1990. Survey on a shoestring: A manual for small-scale language surveys. Dallas: Summer Institute of Linguistics and University of Texas at Arlington.
6. Grimes, Joseph E. 1995. Language survey reference guide. Dallas: Summer Institute of Linguistics, Inc.
7. Kluge, Angela. 2006. RTT retelling method: An alternative approach to intelligibility testing. SIL International. Dallas, TX.
8. Chafe W. (ed.). The pear stories: Cognitive, cultural, and linguistic aspects of narrative production. – Norwood: Ablex, 1980.

References

1. Mutalov, R.O. (2002) Glagol darginskogo yazyka. [Verb of the Dargin language]. Mahachkala: Izdatel'sko-poligraficheskij centr DGU. 216 p. (In Russ.)
2. Mutalov, R.O. The classification of the Dargin languages and dialects [online] // Sociolinguistics, 2021. No. 3 (7). Pp.8-25. (In Russ.) DOI: 10.37892/2713-2951-3-7-8-25

[6] Даргинско-русский словарь. 12000 слов и фразеологических выражений. – Махачкала: АЛЕФ (ИП Овчинников), 2017. – 648 с.

[7] Text Encoding Initiative. Режим доступа [URL]: https://tei-c.org/

[8] Каталог ВКР НИУ ВШЭ. Магомедов А.Г. Создание рабочей среды для обработки словарей на примере даргинского словаря. Режим доступа [URL]: https://www.hse.ru/ma/dh/students/diplomas/839720080

3. Malyshev, V., Malysheva, V., Gutz, A., Novaya, I., Panina, A., Yurkova, A., Clifton, J.M., Tiessen, C. (2019). The Sociolinguistic Situation of the Dargwa in Dagestan. SIL Electronic Survey Report 2019-011. Dallas: SIL International. 51 p.
4. Casad, Eugene. 1974. Dialect intelligibility testing. Norman, OK: Summer Institute of Linguistics.
5. Blair, Frank. 1990. Survey on a shoestring: A manual for small-scale language surveys. Dallas: Summer Institute of Linguistics and University of Texas at Arlington.
6. Grimes, Joseph E. 1995. Language survey reference guide. Dallas: Summer Institute of Linguistics, Inc.
7. Kluge, Angela. 2006. RTT retelling method: An alternative approach to intelligibility testing. SIL International. Dallas, TX.
8. Chafe W. (ed.). The pear stories: Cognitive, cultural, and linguistic aspects of narrative production. – Norwood: Ablex, 1980.