# Creating a Spoken Corpus of Yakut-Russian Code-Switching

**Anna Petukhova**
National Research University Higher School of Economics, School of Linguistics
annap2.71828tukhova@gmail.com

**Elena Sokur**
National Research University Higher School of Economics, Linguistic Convergence Laboratory
elena.o.sokur@gmail.com

**Abstract**

This paper outlines a spoken corpus of Yakut-Russian code-switching based on the conversation during a board game session. The audio recordings were transcribed, aligned with the texts in ELAN and annotated according to the original tagging strategy that provides a further investigation of code-switching. The corpus is oriented to language contact research based on code-switching materials. The corpus uses a search platform tsakorpus and is available at the http://lingconlab.ru/cs_yakut/. The paper describes the process of corpus development, including technical implementation, and presents case studies that use the corpus data.

**Keywords:** corpus linguistics, spoken corpora, Russian, Yakut, code-switching
**DOI:** 10.28995/2075-7182-2021-20-1161-1169

# Создание устного якутско-русского корпуса переключения кода

**Петухова Анна Андреевна**
Национальный исследовательский университет «Высшая школа экономики», Школа лингвистики
annap2.71828tukhova@gmail.com

**Сокур Елена Олеговна**
Национальный исследовательский университет «Высшая школа экономики», Лаборатория языковой конвергенции
elena.o.sokur@gmail.com

**Аннотация**

Статья представляет устный корпус переключения кода якутско-русских билингвов, в основу которого легла беседа во время партии настольной игры. Аудиозаписи были расшифрованы, выровнены с текстом в программе ELAN и размечены в соответствии со стратегией, разработанной специально для этого корпуса. Корпус предназначен для изучения языковых контактов на материале переключений кода. Корпус использует поисковую платформу tsakorpus и доступен по адресу http://lingconlab.ru/cs_yakut/. В статье описывается процесс создания корпуса, включая технологическое решение, и сценарии исследований на основе материалов корпуса.

**Ключевые слова:** корпусная лингвистика, устный корпус, русский язык, якутский язык, переключение кода

## 1 Введение

Переключение кода – чередование двух языков в пределах дискурса, когда внутри языкового высказывания говорящий переключается с одного идиома на другой (см. пример 1, переключение с русского на якутский, здесь и далее в квадратных скобках указан говорящий и отметка времени высказывания, слова на якутском языке выделены жирным). Оно является характерной чертой языкового поведения билингвов (Bullock & Toribio 2009).

(1) Что за **уон** **ахсыс**?
**десять** **восьмой**
‘Что за восемнадцатый?’ [говорящий3, 1341.43]

Настоящая статья представляет устный якутско-русский корпус переключения кода, подробно описывает шаги его создания и предлагает возможные сферы его применения – как в изучении переключения кода в частности, так и в исследовании языковых контактов в целом.
В России билингвизм особенно распространен на территории республик, в том числе в Якутии, где якутский язык признан государственным наряду с русским: по данным переписи населения 2010 года (www.gks.ru, дата обращения: 20.05.2020) большинство этнических якутов, а также до половины русских в Республике Саха (Якутии), являются билингвами (Иванова 2013).

Корпусный метод – один из современных и результативных подходов к изучению переключения кода (Lonngren Sampaio 2015: 32). Основными источниками данных для билингвальных корпусов являются элицитация (Lyu et al. 2015), записи спонтанной речи (Poplack 1980, Halmari & Smith 1994, Huddlestone & Nel 2012, Lyu et al. 2015) и тексты из сети Интернет (Biemann et al. 2013, Maharjan et al. 2015). Корпуса, составленные на основе спонтанной речи, имеют большое преимущество перед письменными текстами (например, тексты газет, книг или сообщения из социальной сети “Твиттер”). Переключение кода может быть стигматизировано в билингвальных сообществах (Travis & Cacoullos 2013), и говорящие могут избегать переключений, осознанно отвечая на вопросы исследователя, как это происходит при элицитации. Поэтому корпуса, составленные на основе письменных текстов, могут не содержать высказываний с переключением кода или содержать существенно меньше вхождений этого языкового феномена, особенно если один из двух языков принадлежит этническому меньшинству.

На настоящий момент корпусов переключений кода очень мало, так как создание таких корпусов времязатратно: записи и их расшифровки, а также обучение разметке времязатратны. Также встает проблема этичности обнародования таких языковых материалов – говорящие не всегда соглашаются на их публикацию.

Наиболее близким к представленному в нашей статье корпусу является корпус переключения кодов четырех малых языков: тунгусо-манчжурских (нанайский, ульчинский) и уральских (горно-марийский, мокшанский) (http://web-corpora.net/ruscontact/CS.html), также снабженный специальной разметкой переключения кода и контактно-обусловленных признаков на разных языковых уровнях (Khomchenkova et al. 2019). В основе этого корпуса лежат тексты спонтанной речи, собранные в рамках проекта по документации малых языков.

Якутско-русский корпус переключения кода, представленный в настоящей статье, отличается тем, что это устный корпус, содержащий расшифровки спонтанной речи, записанной специально для исследования переключения кода в результате эксперимента. В нашем корпусе есть разметка случаев переключения кода с пометкой метаданных (подробнее см. в Таблице 1), а также слои, указывающие язык, на который произошло переключение кода. Эти слои выровнены с предложением, что позволяет найти все высказывания определенного типа переключения кода (внутрисентенциального или межсентенциального) с указанием конкретного языка (русский, якутский).

| **Межсентенциальные (intersentential):** | |
|---|---|
| **Лексические** | |
| IDIOM | идиома (idiom) |
| INTJ | междометие (interjection) |
| **Метаязыковые** | |
| QUOTE | цитирование (quote) |

| | |
|---|---|
| **Внутрисентенциальные (intrasentential):** | |
| **Морфологические** | |
| ADJ | прилагательное (adjective) |
| ADV | наречие (adverb) |
| AUX | вспомогательный глагол (auxiliary) |
| MODP | модальная частица (modal particle) |
| N | существительное (noun) |
| QP | вопросительная частица (question particle) |
| V | глагол (verb) |
| **Синтаксические** | |
| ADJP | группа прилагательного (adjective phrase) |
| ADVP | наречная группа (adverb phrase) |
| CONJ | подчиненное придаточное (conjunctive clause) |
| COORD | сочиненное придаточное (coordinate clause) |
| IC | независимая клауза (independent clause) |
| NP | именная группа (noun phrase) |
| PP | предложная группа (prepositional phrase) |
| VP | глагольная группа (verb phrase) |
| **Встречаются в обоих видах (both):** | |
| **Морфологические** | |
| DE | дискурсивная единица (discourse entity) |
| P | предлог (preposition) |
| Q | вопрос (question) |
| REP | повтор фразы (utterance repetition) |
| **Язык, на который происходит переключение кода:** | |
| r | русский (Russian) |
| y | якутский (Yakut) |
| e | английский (English) |

| **Другие обозначения:** | |
|---|---|
| **Метаязыковые** | |
| <> | игровой термин |
| {} | высказывание, произнесенное вне беседы (например, во время телефонного разговора) |
| [INDCPH] | неразборчиво |
| **Просодические** | |
| = | обрыв высказывания |

Таблица 1: Список обозначений

## 2 Эксперимент

Записи спонтанной речи, на которые опираются исследователи в своих работах, могут быть собраны как во время интервью (билингвы разговаривают с билингвами: Poplack 1980, Lyu et al. 2015), так и во время эксперимента – искусственно созданных условий (Halmari, Smith 1994; Huddlestone, Nel 2012). Для нашего исследования был выбран второй метод – эксперимент, в рамках которого участники играли в настольную игру. Игра удобна тем, что требует минимальной подготовки, может длиться продолжительное время, а также вовлекает всех участников в разговор в равной степени, что нельзя проконтролировать в реальной беседе. Информанты играли в настольную игру «Монополия», предусматривающую активную коммуникацию всех участников и не имеющую строгой продолжительности (средняя длительность игры – 2 часа). «Монополия» – экономическая стратегическая игра, цель которой – добиться банкротства других игроков, используя стартовый капитал. Ход игры состоит в том, что игроки по очереди бросают кубики и совершают ходы на игровом поле. Каждой клетке поля соответствует актив (город, транспортная компания, коммунальное предприятие) или событие (шанс, общественная казна, бесплатная парковка, отправление в тюрьму). Оказавшись на клетке поля, игрок либо покупает актив, либо платит владельцу арендную плату.

Мы попросили говорящих играть в настольную игру, при этом не объясняя им, что смысл эксперимента – собрать образцы речи с переключением кода. В ходе эксперимента было получено 148 минут спонтанной речи в рамках одной сессии игры (см. общую статистику корпуса в таблице 2). Во время игры автор эксперимента находилась в другом конце комнаты и не участвовала в беседе. Разговор информантов записывался в тихой комнате на диктофон, записи были сохранены в формате .wav.

Информанты дали согласие на использование записей их голоса в исследовательских целях. После проведения эксперимента участникам была разъяснена цель исследования.

| | |
|---|---|
| Минуты | 148 |
| ЭДЕ (элементарная дискурсивная единица) | 5 870 |
| Токены | 19 325 |
| Переключения кода | 7 822 |
| из них межсентенциальные: | 5 870 |
| из них внутрисентенциальные: | 1 952 |

Таблица 2: Статистика корпуса

## 3 Участники

В эксперименте приняли участие четверо якутско-русских билингвов: трое мужчин и одна женщина. Возраст участников – от 18 до 19 лет. На момент проведения эксперимента все информанты проживали в городе Якутске от 4 до 19 лет, являлись студентами университета и были знакомы друг с другом от двух до трех лет, что способствовало созданию непринужденной беседы во время эксперимента. Для минимизации количества факторов, влияющих на переключение кода, были отобраны участники с максимально схожим языковым опытом: одного возраста, с родным якутским языком и хорошим владением русским языком.

## 4 Размер корпуса и стратегия разметки

Статистика корпуса приведена в таблице 2. Элементарной дискурсивной единицей мы считаем клаузы, обладающие интонационной целостностью (Подлесская & Кибрик 2007).

Стратегия разметки в исследованиях языковых контактов целиком определяется задачами научно-исследовательской работы и, следовательно, является уникальной для каждого исследования. Различают два типа переключения кода – внутрисентенциальный (внутри высказывания, см. пример 2) и межсентенциальный (между высказываниями, см. пример 3).

(2) Десять миллионов эбиллээн биэрэр.
добавляет AUX
‘Десять миллионов добавляет (тебе)’. [говорящий3, 5301.74]

(3а) Да, мое что-то украли вообще! [говорящий2, 2922.62]

(3б) Ол иһин аччаан хаалбыта э.
тот CONJ уменьшилось AUX INTJ
‘Поэтому уменьшилось, э’. [говорящий2, 2925.52]

Наше исследование опирается на процедуру аннотации, предложенную Шаной Поплак, где наряду со случаями переключения кода отдельно отмечаются синтаксические категории составляющей на границе переключения кода (Poplack 1980). Для каждого говорящего были размечены 6 слоев (разметка проводилась в программе ELAN): элементарная дискурсивная единица (ЭДЕ), слова на русском языке, слова на якутском языке, перевод слов на якутском языке, синтаксическая категория составляющей, на границе с которой происходит переключение кода (см. табл. 2), а также внутрисентенциальные и межсентенциальные переключения кода с указанием языка, на который произошло переключение кода (см. рис. 1). Отдельно были отмечены высказывания, произнесенные вне беседы (например, во время телефонного разговора одного из информантов за игральным полем), а также высказывания, представляющие собой специальные термины игры, и чтение вслух правил игры и игральных карточек. В примере 4 слово дубль, являющееся внутриигровым понятием, помещено в треугольные скобки и входит в наречную группу со словом эмиэ, поскольку язык игры был изначально задан предоставленной русскоязычной версией «Монополии», и говорящие были ограничены в выборе языковых единиц, касающихся процесса игры.

(4) Десять,<дубль> эмиэ.
опять
‘Десять, опять <дубль>’. [говорящий1, 6119.02]

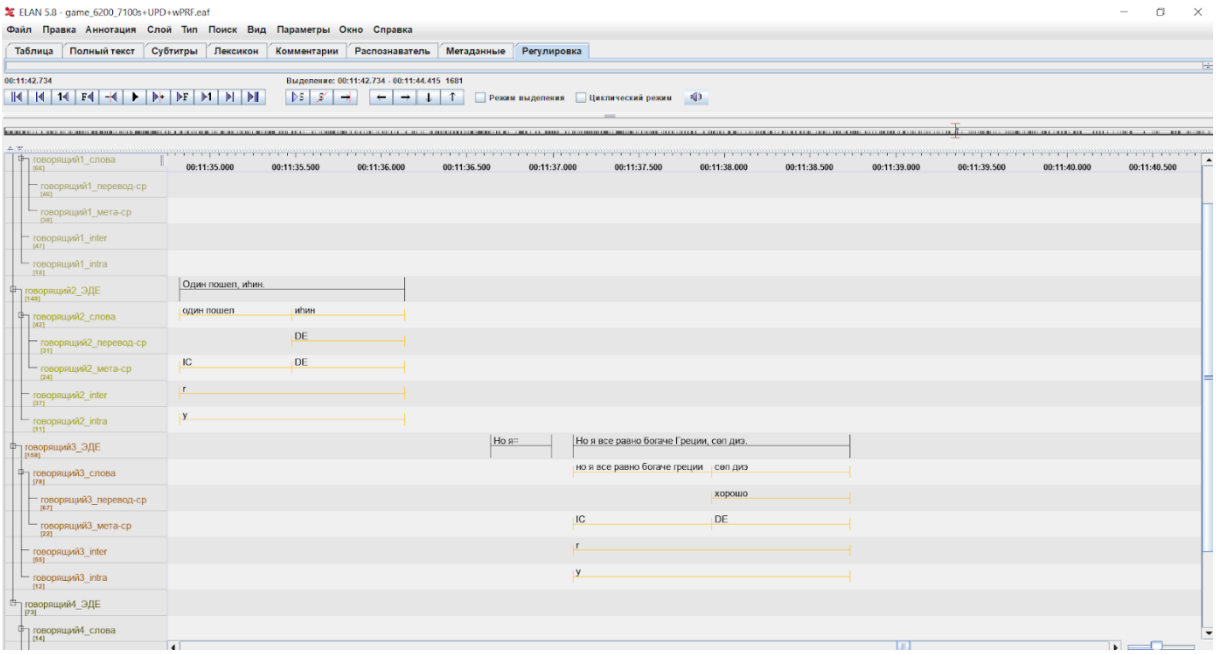


Рисунок 1: Пример из ELAN

Высказывания были выровнены с аудиозаписями в программе ELAN, что позволяет пользователю опираться на первичные языковые данные, самому прослушивая расшифрованную речь билингвов.

Для облегчения разметки полная запись беседы была разбита на 10 отрезков длиной 13-15 минут, расшифровки которых легли в основу корпуса.

## 5 Возможности поиска

Задача состояла в том, чтобы создать корпус, интерфейс которого был бы удобен в использовании для изучения контактного явления переключения кода. Для этого была использована поисковая платформа tsakorpus (https://bitbucket.org/tsakorpus/tsakorpus/src/master/). Tsakorpus – это лингвистическая поисковая платформа, которая использует поисковую систему elasticsearch для хранения и поиска данных. Платформа поддерживает корпуса с морфологической разметкой, глоссированием, выравненными со звуком текстовыми данными и проч. Платформа также предоставляет удобный поисковый интерфейс с возможностями поиска по метаданным текстов и носителей, выбора подкорпуса, поиска по нескольким словам и точному запросу, использования регулярных выражений. Такая платформа подходит для корпусов, выровненных со звуком в программе ELAN или Praat.

Поиск по корпусу осуществляется по всем слоям в трёх уровнях разметки: якутский, межсентенциальный (переключение кода между высказываниями), внутрисентенциальный (переключение кода внутри высказывания). При поиске по якутскому языку можно искать по точному вхождению (словоформе) внутри ЭДЕ, по точному вхождению в русском переводе ЭДЕ и по тэгам (о них см. таблицу 1 “Список обозначений”). Например, чтобы найти все слова на букву “к”, которые имеют категорию *N* (существительное), надо в поле “Слово” ввести *к**, а в поле “Тэг” ввести *N*. На рис. 2 представлено, как выглядит такой запрос и его поисковая выдача в веб-интерфейсе. Красным цветом выделена ЭДЕ, которая имеет синтаксическую категорию *N* и в которой есть слово, начинающееся на *к*.

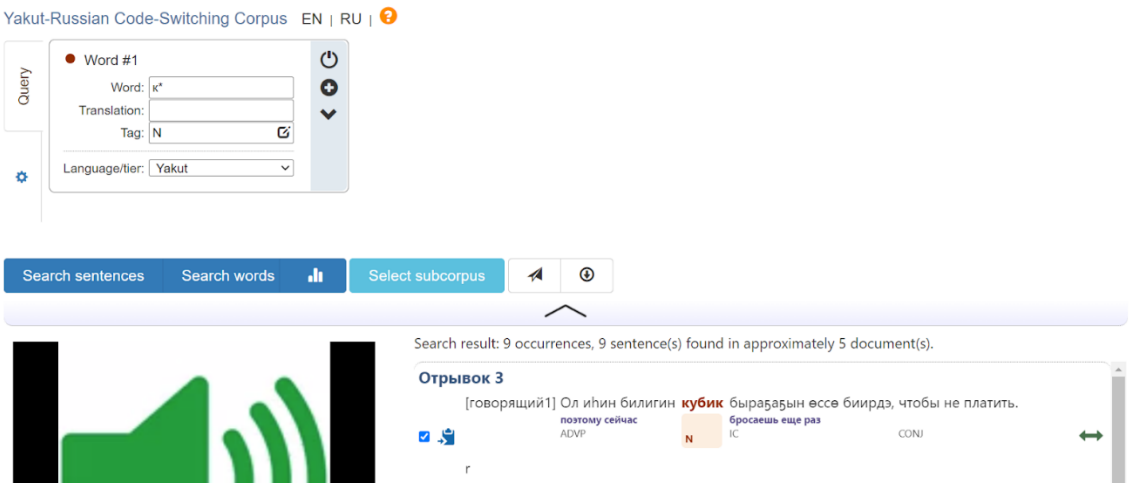


Рисунок 2: Пример поискового ввода и выдачи

При поиске по межсентенциальному и внутрисентенциальному уровням действительно только первое поле поиска “Слово”, куда вводятся три сентенциальных тэга: *y* (якутский), *r* (русский), *e* (английский). Например, чтобы найти все предложения с переключением кода с якутского, нужно выбрать слой Межсентенциальный и ввести y в поле “Слово”.

Поисковая выдача устроена таким образом, что для каждого предложения видно, какие в нём есть ЭДЕ, как они переводятся (если ЭДЕ на якутском) и какой у них тэг, а также с какого языка произошло переключение кода (при условии, что оно произошло).

Каждое предложение выровнено со звуком. Чтобы прослушать предложение, надо нажать на верхний уровень разметки.

В настоящее время корпус доступен онлайн по ссылке http://lingconlab.ru/cs_yakut/.

## 6 Использование корпуса

Первая работа, в которой был использован материал корпуса – описание стратегий переключения кода якутско-русских билингвов (Петухова 2020). Количественный анализ данных корпуса показал, что говорящие чаще переходили на русский язык между высказывания (межсентенциальные переключения), а внутри высказывания – на якутский (внутрисентенциальные перключения). Наиболее распространенными синтаксическими категориями на границе с переключением кода оказались независимые клаузы и дискурсивные единицы – наиболее подвижные составляющие, которые можно вставить в любое место высказывания, не нарушив при этом правил грамматики обоих языков. В результате качественного анализа было выявлено, что в пределах беседы переключение кода происходило в моменты изменения темы, введения персонального комментария к игровой ситуации и провала коммуникативной стратегии, что характеризует переключение кода, ориентированное на дискурс. Выводы свидетельствуют в пользу утверждений Ш. Поплак и П. Ауэра о чувствительности переключения кода к дискурсу и грамматике обоих языков (Auer 1995, Poplack 1980).

Далее количественные данные корпуса вместе с результатами социолингвистической анкеты и тестами для определения уровня владения языками были использованы для проверки гипотезы о корреляции типов переключения кода с уровнем билингвизма говорящих. Существует предположение, что билингвы, владеющие обоими языками одинаково хорошо, будут предпочитать чаще переключать код внутри высказывания, чем между высказываниями, так как такой тип переключения кода требует достаточных знаний грамматики обоих языков. Впервые эта гипотеза была выдвинута Шаной Поплак и в последующем поддержана в исследованиях с такими языковыми парами, как английский-испанский (Poplack 1980, Toribio 2001, Zentella 1997, Gollan & Ferreira 2009, Aguirre 1985), английский-турецкий (Koban 2013) , английский-севернокитайский (Yow et al. 2018). Для пары типологически неродственных языков якутский-русский эта гипотеза не была подтверждена: данные корпуса показали, что частота внутрисентенциальных переключений кода, которые считаются отличительной чертой сбалансированных билингвов, не зависят от субъективной и объективной оценки уровня билингвизма говорящих.

## 7 Заключение

В статье был представлен новый тип устного корпуса – якутско-русский корпус переключения кода, описан процесс его создания и приведены сценарии исследований с использованием корпусного материала.

Якутско-русский корпус переключения кода вносит вклад в корпусные исследования русского языка и его влияния на другие языки России. Нами была разработана система разметки, которая может быть использована при создании новых корпусов переключения кода. Такая разметка вместе с поисковым интерфейсом может быть использована для проведения квантитативных исследований контактных явлений.

## 8 Список сокращений

AUX – вспомогательный глагол
CONJ – подчинительный союз
INTJ – междометие

## Благодарности

Мы благодарим информантов за их терпение и энтузиазм.

Статья и корпус созданы в рамках Программы фундаментальных исследований Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ).

## Литература

[1] Aguirre Jr. A. (1985), An experimental study of code alternation, International Journal of the Sociology of Language. Vol. 53, pp. 59-82.
[2] Auer P. (1995), The pragmatics of code-switching: A sequential approach, One Speaker, Two Languages: Cross-Disciplinary Perspectives on Code-Switching, Cambridge University Press, Cambridge, pp. 115-135.
[3] Biemann C., Bildhauer F., Evert S., Goldhahn D., Quasthoff U., Schäfer R., Zesch T. (2013), Scalable Construction of High-Quality Web Corpora, Journal for Language Technology and Computational Linguistics, Vol. 28, pp. 23-59.
[4] Bullock B. E., Toribio A. J. (2009), Themes in the study of code-switching, The Cambridge Handbook of Linguistic Code-switching, Cambridge University Press, Cambridge, pp. 1-17.
[5] Gollan T. H., Ferreira V. S. (2009), Should I stay or should I switch? A cost–benefit analysis of voluntary language switching in young and aging bilinguals, Journal of Experimental Psychology: Learning, Memory, and Cognition, Vol. 35, pp. 640-665.
[6] Halmari H., Smith W. (1994), Code-switching and register shift: Evidence from Finnish-English child bilingual conversation, Journal of Pragmatics, Vol. 21, pp. 427-445.
[7] Huddlestone K., Nel J. (2012), Analysing Afrikaans-English bilingual children's conversational code switching, Stellenbosch Papers in Linguistics, Vol. 21, pp. 29-53.
[8] Ivanova N. I. (2013), Russian Language Status in the Republic of Sakha (Yakutia) [Status russkogo yazy`ka v Respublika Saxa (Yakutiya)], Bulletin of NWFU [Vestnik SVFU], Vol. 10, pp. 58-63.
[9] Khomchenkova I. A., Pleshak P. S., Stoynova N. M., (2019), The Corpus of Contact-influenced Russian of Northern Siberia and the Russian Far East, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2019” [Komp’yuternaya Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii “Dialog 2019”], pp. 276-287.
[10] Koban D. (2013), Intra-sentential and inter-sentential code-switching in Turkish-English bilinguals in New York City, US, Procedia-Social and Behavioral Sciences, Vol. 70, pp. 1174-1179.
[11] Lonngren Sampaio C. A. (2015), The investigation of code-switching in a computerised corpus of child bilingual language.
[12] Lyu D., Tan T., Chng E., Li H. (2015), Mandarin–English code-switching speech corpus in South-East Asia: SEAME, Lang Resources & Evaluation, Vol. 49, pp. 581-600.
[13] Maharjan S., Blair E., Bethard S., Solorio T. (2015), Developing language-tagged corpora for code-switching tweets, Proceedings of The 9th Linguistic Annotation Workshop, Denver, pp. 72-84.

[14] Petukhova A. A., (2020), Corpus Research on Code-Switching in Yakut L1 Bilinguals [Korpusnoe issledovanie pereklyucheniya koda v besede bilingvov s rodny`m yakutskim yazy`kom], The Second Conference on Uralic, Altaic and Paleo-Siberian Languages. Abstracts of Internartional Sciemtific Conference [Вторая конференция по уральским, алтайским и палеоазиатским языкам. Тезисы докладов международной научной конференции], Saint-Petersburg, pp. 71-73.
[15] Podlesskaya V. I., Kibrik A. A. (2007), Подлесская, Кибрик 2007 – В. И. Подлесская, А. А. Кибрик. Speaker's Self-Correction and Other Types of Speech Failures as Annotation Objects in Spoken Corpora [Samoispravleniya govoryashhego i drugie tipy` rechevy`x sboev kak ob``ekt annotirovaniya v korpusax ustnoj rechi], Scientific and Technological Information [Nauchno-texnicheskaya informaciya], Vol. 2, pp. 2–23.
[16] Poplack S. (1980), Sometimes I'll start a sentence in Spanish Y TERMINO EN ESPANOL: toward a typology of code-switching, Linguistics, Vol. 18, pp. 581-618
[17] Travis C. E., Cacoullos R. T. (2013), Making voices count: Corpus compilation in bilingual communities, Australian Journal of Linguistics, Vol. 33, pp. 170-194.
[18] Yow W. Q., Tan J. S. H., Flynn S. (2018), Code-switching as a marker of linguistic competence in bilingual children, Bilingualism: Language and Cognition, Vol. 21, pp. 1075-1090.
[19] Zentella A. C. (1997), Growing up bilingual, Blackwell, Malden.