Переключение кодов в устных корпусах малых языков

Н. М. Стойнова, stoynova@yandex.ru

Круглый стол «Корпусные методы в исследовании языковых контактов»

МЕЖДУНАРОДНАЯ НАУЧНАЯ КОНФЕРЕНЦИЯ, ПОСВЯЩЕННАЯ
20-ЛЕТНЕМУ ЮБИЛЕЮ НАЦИОНАЛЬНОГО КОРПУСА РУССКОГО ЯЗЫКА

Москва, ИРЯ РАН, 20-21.12.2024

Устные корпуса малых языков: современный тренд

→ Все больше и больше устных корпусов малых и вымирающих языков

→ Золотой стандарт современного проекта по языковой документации

- отглоссированная коллекция текстов
- синхронизированная со аудио- и часто с видеозаписью
- часто с удобным поисковым онлайн-интерфейсом
- часто в открытом доступе

→ Малые языки могут опережать по “ресурсности” большие и средние

Устные корпуса малых языков: современный тренд

→ Например, коллекции текстов на малых исчезающих языках мира

- архив ELAR (https://www.elararchive.org/)

→ Например, для языков России

- корпуса Лаборатории исследования и сохранения малых языков ИЯз РАН (https://corpora.iling-ran.ru/)
- корпуса Лаборатории языковой конвергенции НИУ ВШЭ (https://lingconlab.ru/resources_ru.html#m1)

Исследование переключения кодов на материале устных корпусов малых языков

→ Открывает принципиально новые возможности для исследования переключения кодов

→ Тексты с переключением кодов (*т.е. с фрагментами, иногда большими, на лингва-франка и/или других языках многоязычного региона*)

- часто составляют значительную часть полевой коллекции
- особенно для вымирающих языков

→ При этом:

- задачи создать специальный ресурс для исследования переключения кодов в проектах по документации обычно не ставится

В этом докладе

- Корпуса малых языков России
- Переключение кодов с русским языком

- Удобства и неудобства корпусов для исследования переключения кодов
- Какие характеристики корпусов релевантны для исследования переключения кодов?
- Разметка переключения кодов: нужна ли и какая?

Удобства

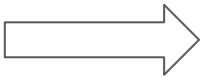
→ Привлечение данных за пределами больших языков

- для них переключение кодов значительно лучше исследовано (напр., английский-испанский)

→ **Возможность работать с отглоссированными текстами**

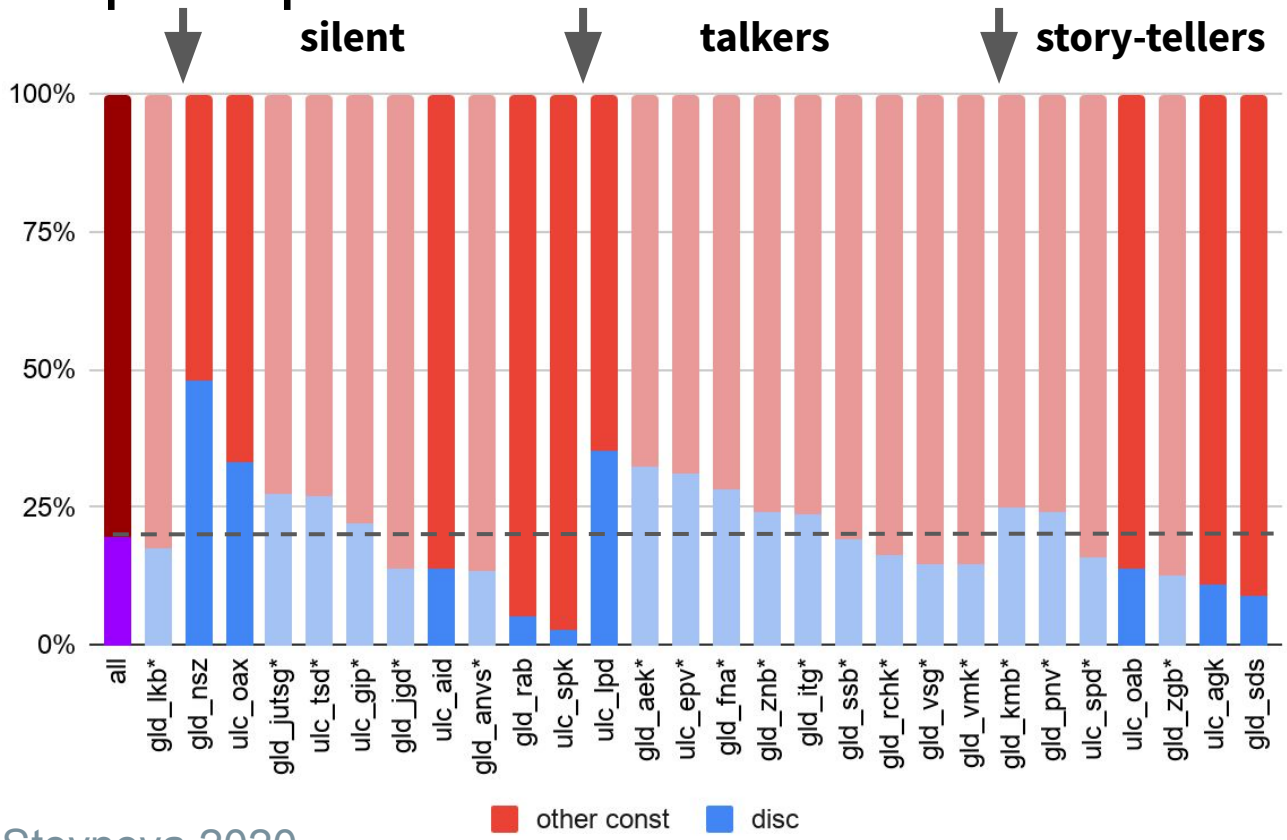
- коллекции текстов, собираемых специально для исследования переключения кодов, как правило, не содержат морфологической разметки

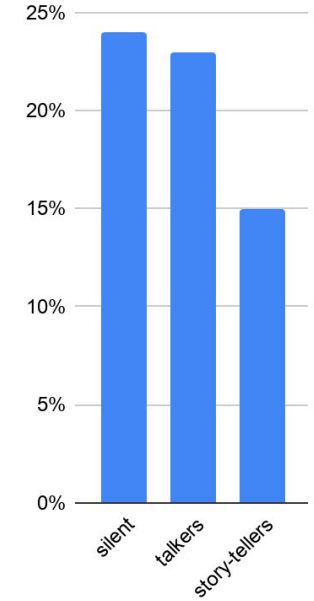
вопросы типа:
Что частотнее – переключенные ИГ или дискурсивные маркеры?



вопросы типа:
Какова доля переключенных ИГ из всех ИГ? Доля переключенных дискурсивных маркеров из всех?

Пример





Процент переключенных дискурсивных маркеров (от всех переключений) у разных носителей нанайского и ульчского

Stoynova 2020

Удобства

→ Синхронизация со звуком

→ В целом ориентация на максимально точную запись устной речи

→ Относительная сопоставимость корпусов в смысле жанров и типов записанных текстов

- *нарративы: рассказы о себе, рассказы о прошлом…*

Неудобства

→ **Малый объем коллекций**

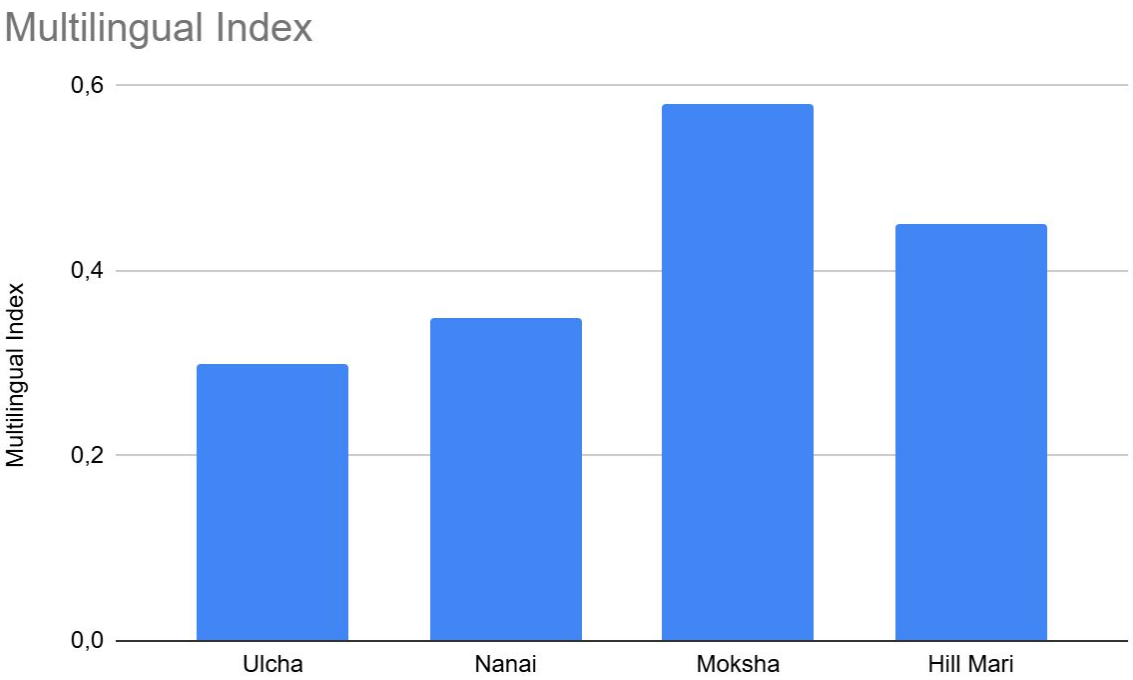
→ **Не сбалансированы по социолингвистическим параметрам**

- NB корпуса вымирающих языков в выгодном положении: охвачены все (оставшиеся) носители

→ **Сознательный отбор текстов с минимумом переключения кодов на этапе создания корпуса**

- NB корпуса вымирающих языков в выгодном положении: в корпус входит “все, что есть”
- NB вопрос о сопоставимости корпусов языков на разных стадиях языкового сдвига

Пример



→ Объем русского материала в корпусе говорит о стадии языкового сдвига или о составе корпуса?

Dyachkov et al. 2020

Неудобства

→ **Очевидная искусственность самой ситуации записи текста**

- нарративы по инструкции лингвиста “расскажите на своем языке”
- NB лингвисту-носителю русского

- обычно в фокусе внимания “спонтанное переключение кодов”
- спонтанно тот же человек меньше бы контролировал переключения на русский
- для ситуации языкового сдвига: спонтанно тот же человек рассказал бы эту историю по-русски
- NB корпуса вымирающих языков в выгодном положении: рассказы для лингвиста – единственная форма функционирования языка

Что релевантно для исследования переключения кодов?

→ Подробность метаданных

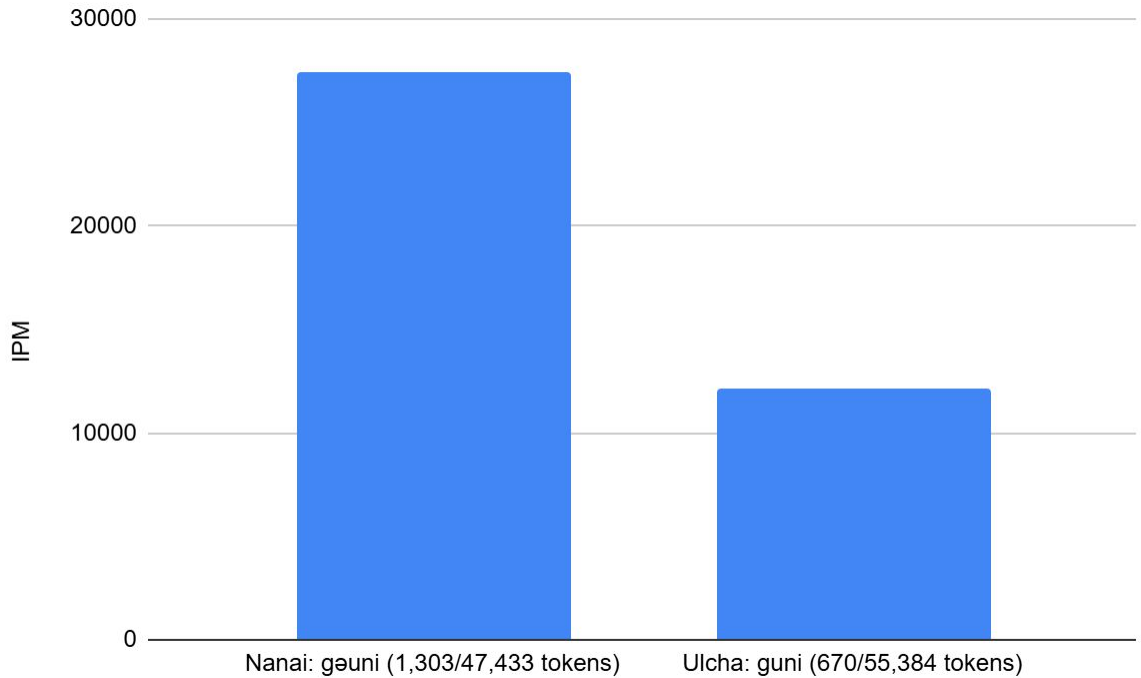
- наличие / отсутствие информации о носителях и другой метаинформации (напр., спонтанный текст? чтение по бумажке?)

→ **Границы ЭДЕ / предложений / клауз**

→ **Границы словоформ**

- нормирование: X русских словоформ на 1,000 словоформ
- нормирование: X переключений на 1,000 предложений

Пример



Нанайская коллекция

- частицы-клитики (напр., *-gəuni*) размечены как часть слова

Ульчская коллекция

- частицы-клитики (напр., *=guni*) размечены как отдельные слова

Stoynova 2021

Оформление иноязычных фрагментов

→ Объема словоформы / меньше словоформы

- расшифрованы, могут быть / не быть отглоссированы
- иногда, но редко как-то специально помечены

→ Объема меньше предложения

- обычно расшифрованы, оформлены бывают по-разному

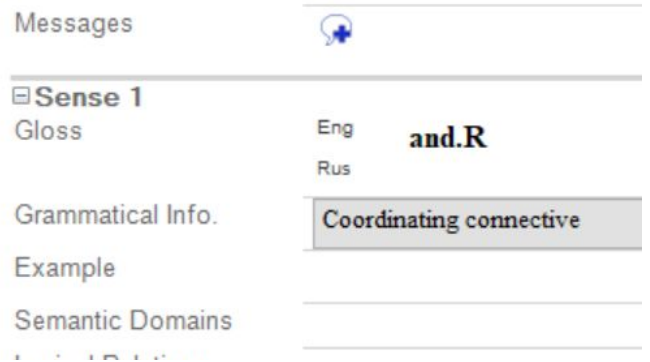
→ Объема предложения и больше

- могут быть не расшифрованы, иногда можно / иногда нельзя найти

Фрагменты объема словоформы или меньше

→ Очень удобно, когда иноязычная основа помечена в глоссах

belyj griby	belyj griby	porcini.R	Noun
benzin	benzin	gasoline.R	Noun
bəŋɑ:j	bəŋɑ:j	granddaughter	Noun
bəŋnaldə	bəŋnaldə	fool.around	Verb (1v)
berek	berek	riverbank.R	Noun (3n)
beresta	beresta	birchbark.R	Noun (1n)



13	**Word**	konečno	usilkan	ɟəgdəŋatinin			
	Lex. Gloss	of.course.R	birchbark.house	burn[intr]	-deont	-3sg	

же
ptl.R

Free Eng Of course a birchbark house will burn, it's bark, right.
Rus Конечно домик из бересты сгорит, береста же.

негидальский, Корпус Б. Пакендорф и Н. Араловой (https://elar.soas.ac.uk/Collection/MPI1041287)

Фрагменты объема словоформы или меньше

→ **Возможная непоследовательность**

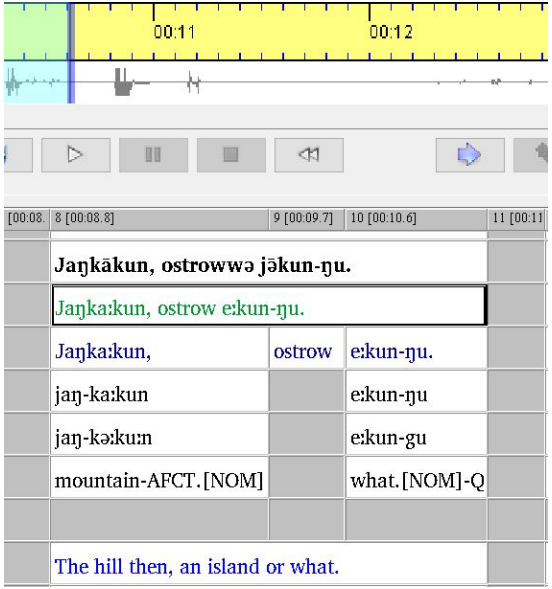
Например

- русский союз *kogda* попал в словарь, набран латиницей, отглоссирован
- русский союз *если* не попал в словарь, набран кириллицей

Например

- существительные с падежными маркерами попали в словарь и отглоссированы
- существительные в номинативе отглоссированы / не отглоссированы

Фрагменты объема словоформы или меньше

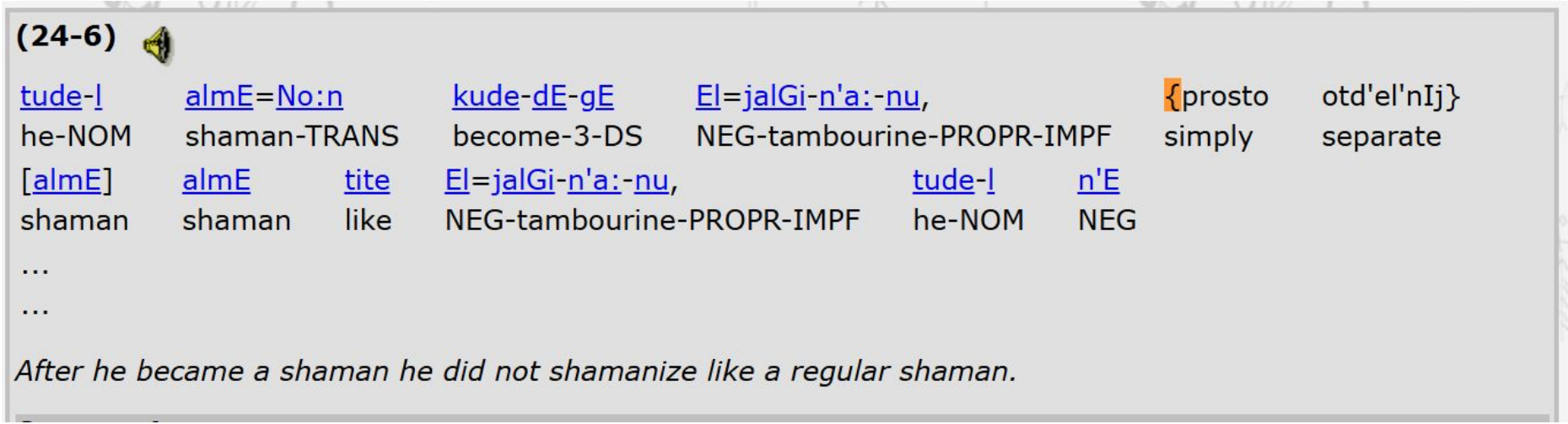


эвенкийский, INEL
(https://inel.corpora.uni-hamburg.de/portal/corpora/evenki/)

	0 [00:01.4	1 [00:02.1]	2 [00:02.7]	3 [00:03.3]	4 [00:03.9]	5 [00:04.5]	6 [00:05.1]
ts-YUK	Tikon dundədu bid'ənoːm aŋ (aŋ-) ostrawdu hakuwsid'anaːrə.						
tx-YUK	Tikon	dundədu	bid'ənoːm	aŋ	(aŋ-)	ostrawdu	hakuwsid'anaːrə.
mb-YUK	tikon	dundə-du	bi-d'ə-noː-m	aŋ	aŋ	ostraw-du	haku-w-si-d'a-naː-rə
mp-YUK	tikəːn	dundə-duː	bi-d'ə-nəː-m	aŋ	aŋ	ostraw-duː	hakuː-wu-sin-d'ə-nəː-rə
ge-YUK	so	earth-DAT/LOC	be-IPFV-PROB-1SG	exactly.this.[NOM]	exactly.this	island-DAT/LOC	encircle-PASS-DUR-IPFV-PROB-AOR.[3PL
BOR-YUK						RUS:core	
fe-YUK	So I am living on the earth, they are probably encircled on this island (?).						

Многословные фрагменты

→ Очень удобно, когда специально помечены



(24-6)

tude-l	almE=No:n	kude-dE-gE	El=jalGi-n'a:-nu,	{prosto	otd'el'nIj}
he-NOM	shaman-TRANS	become-3-DS	NEG-tambourine-PROPR-IMPF	simply	separate

[almE]	almE	tite	El=jalGi-n'a:-nu,	tude-l	n'E
shaman	shaman	like	NEG-tambourine-PROPR-IMPF	he-NOM	NEG

…

…

After he became a shaman he did not shamanize like a regular shaman.

колымский юкагирский, https://typo.uni-konstanz.de/yukaghir/start1.html

Многословные фрагменты

→ Русскоязычные фрагменты кириллицей:

спорное решение, но для поиска переключений очень удобно

ульчский, личная коллекция

Многословные фрагменты меньше предложения

→ **Возможная непоследовательность**

- многословный фрагмент (не отглоссирован) или несколько однословных подряд (с глоссами)?

		12 [00:16.4]	13 [00:17.	14 [00:17.6]	15 [00:18.3]	16 [00:18.9]
ts-XUK						
tx-XUK		kak	((…))	vnuk	kak	bi.
mb-XUK		kak		vnuk	kak	bi
mp-XUK		kak		vnuk	kak	bi
ge-XUK		how		grandson.[NOM]	how	IRREAL
gg-XUK		wie		Enkel.[NOM]	wie	IRREAL
gr-XUK	Z	как		внук.[NOM]	как	IRREAL
mc-XUK		adv		n.[n:case]	adv	ptcl
ps-XUK						

эвенкийский, INEL
(https://inel.corpora.uni-hamburg.de/portal/corpora/evenki/)

ts-XUK		On gund'em, kətə ((…)) tar hutəkar ((…)) əwəd'it kak ((…)) vnuk kak bi.												
tx-XUK		On	gund'em,	kətə	((…))	tar	hutəkar	((…))	əwəd'it	kak	((…))	vnuk	kak	bi.
mb-XUK		on	gun-d'e-m	kətə		tar	hutə-ka-r		əwəd'i-t					
mp-XUK		oni	gu:n-d'ə:-m	kətə		tar	hutə-kə:n-l		əwədi-t					
ge-XUK		how	say-FUT.IMM-1SG	many.[NOM]		that	child-DIM1-PL.[NOM]		Evenki-ADVZ1					

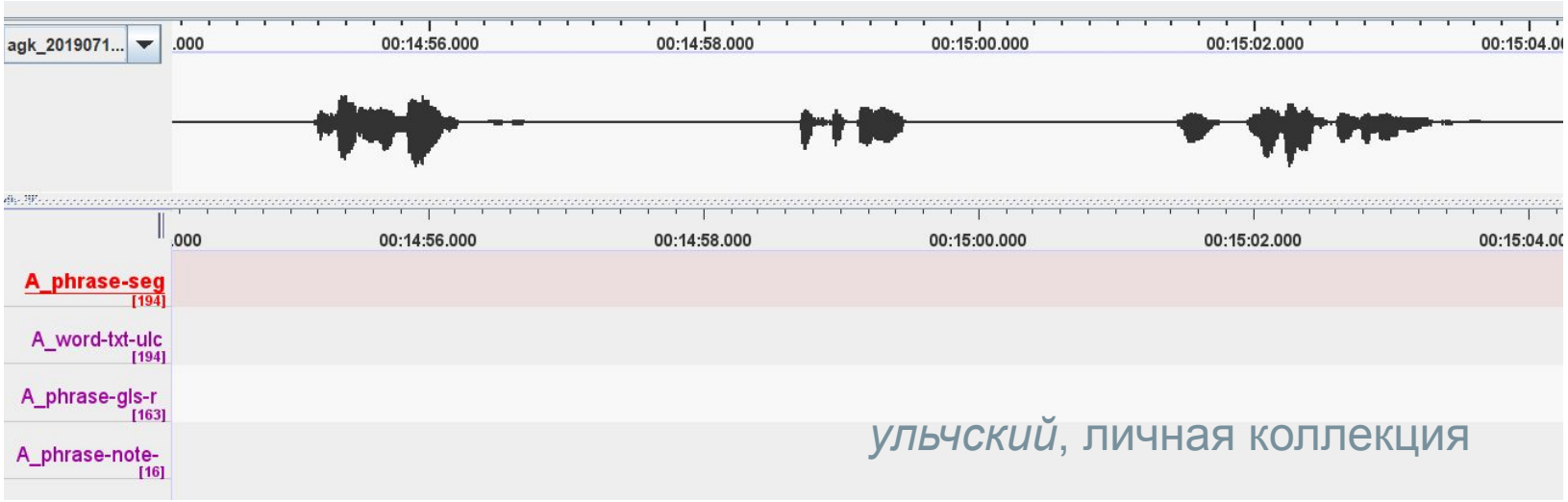
Фрагменты объема предложения и больше

→ Расшифрованы и специально помечены – идеально

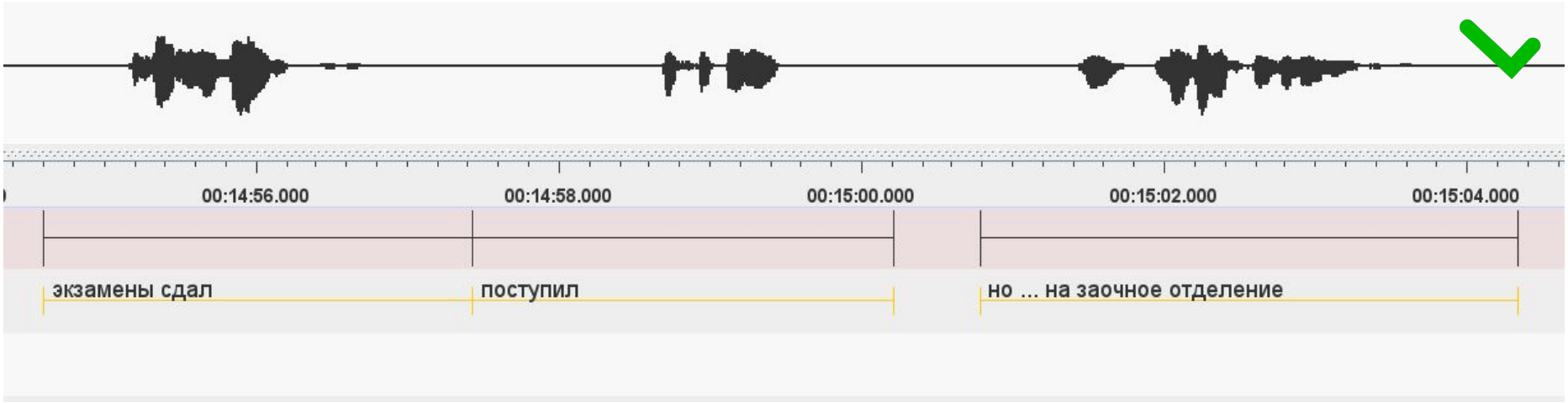
→ Расшифрованы и как-то отличаются от других (например, кириллица / нет перевода / нет глосс / …) – удобно

→ Не расшифрованы, но специально помечены (например, {...}) – нормально

→ Не расшифрованы и никак не ищутся – неудобно



ульчский, личная коллекция



Разметка переключения кодов

→ Проект РФФИ “Переключение кодов в речи русскоговорящих носителей малых языков России” в ИРЯ РАН

- подробная разметка, не интегрирована в корпуса

→ Проект INEL по языкам Сибири университета Гамбурга

- не подробная разметка, интегрирована в корпуса

Разметка переключения кодов в проекте РФФИ

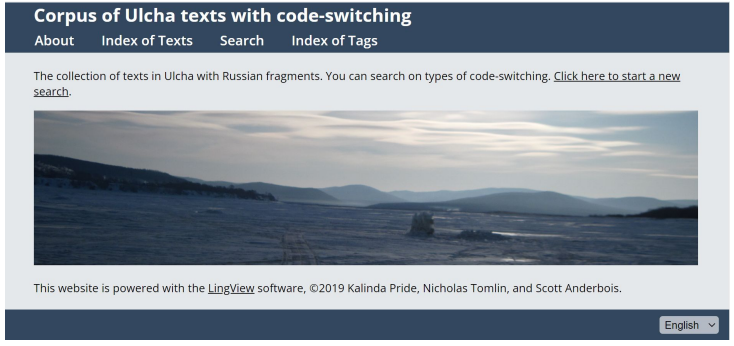
Проект РФФИ “Переключение кодов в речи русскоговорящих носителей малых языков России” (2018–2019)

- http://web-corpora.net/ruscontact/CS.html
- В. Дьячков, П. Плешак, Н. Стойнова, И. Хомченкова
- единообразная подробная разметка переключения кодов в четырех полевых коллекциях

Разметка переключения кодов в проекте РФФИ

Проект РФФИ “Переключение кодов в речи русскоговорящих носителей малых языков России” в ИРЯ РАН (2018–2019)

- не интегрирована в соответствующие корпуса
- размечены все переключения:
  - для части коллекции (горномарийский, мокшанский)
  - для всей коллекции (нанайский, ульчский)
- доступны в виде коллекции в ELAN, для ульчского также онлайн-поиск



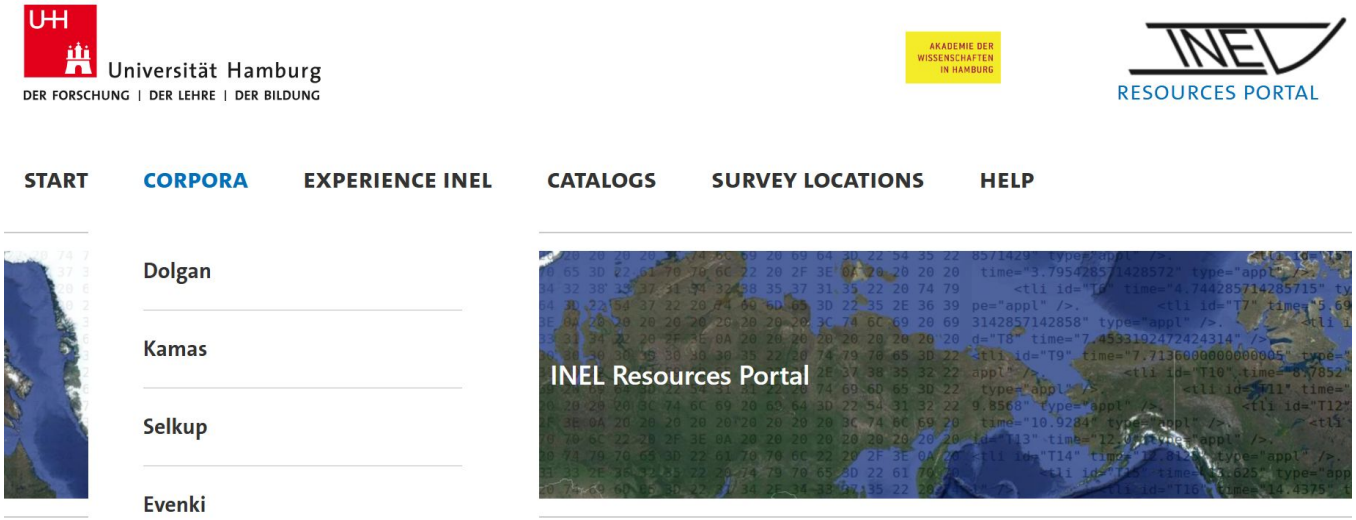
Разметка переключения кодов в проекте РФФИ

Разметка:

- объем (внутрисловное, однословное, многословное, предложение и больше)
- синтаксический тип (разные типы составляющих, не-составляющая)
- дополнительные пометы (хезитация, фальстарт, метаинформация и т.п.)

Разметка переключения кодов в проекте INEL

https://www.slm.uni-hamburg.de/inel/



→ 18-летний продолжающийся проект по созданию корпусов языков Сибири в университете Гамбурга (Grammars, Corpora and Language Technology for **I**ndigenous **N**orthern **E**urasian **L**anguages)

→ Под рук. Беаты Вагнер-Надь

Разметка переключения кодов в проекте INEL

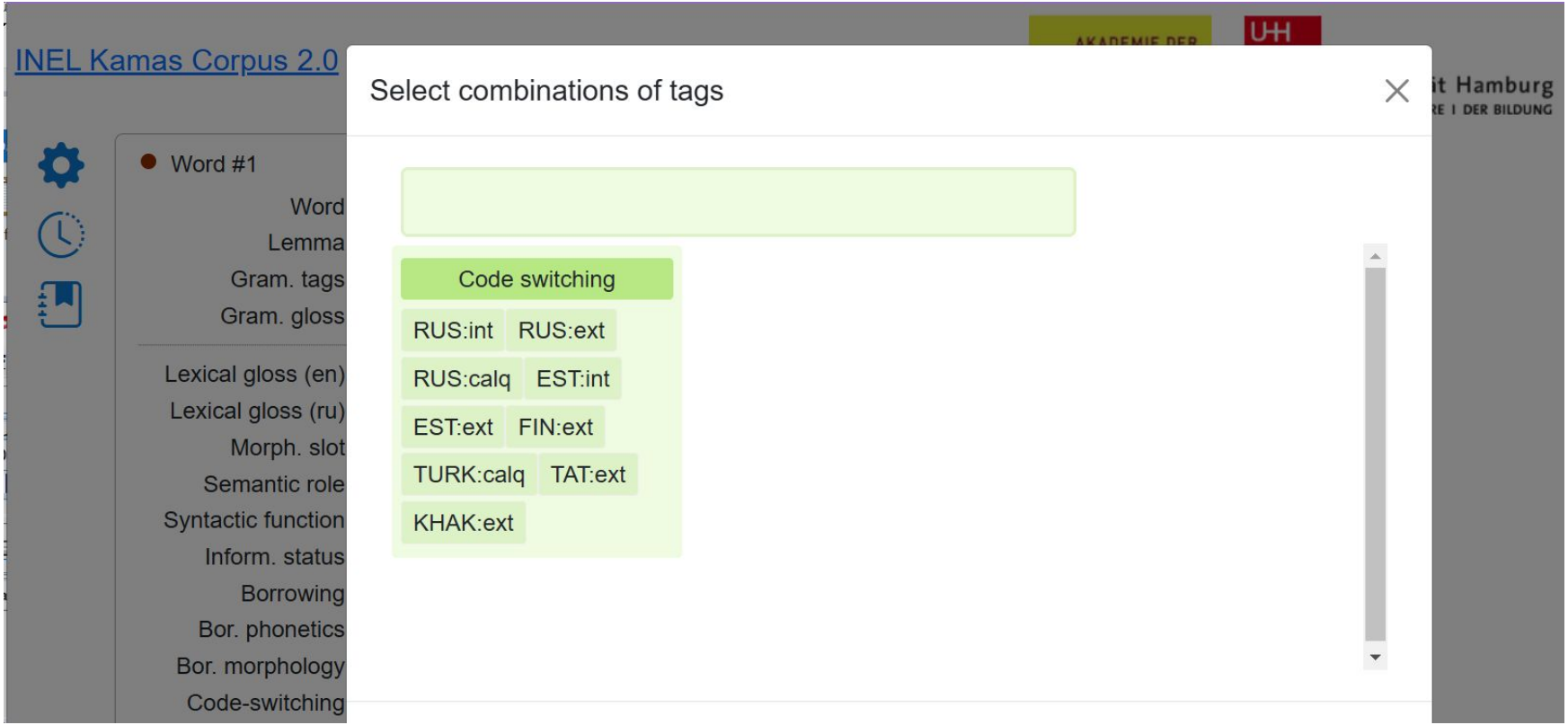
→ Разметка переключения кодов и заимствований – одна из составляющих проекта

→ Разметка заимствований – во всех корпусах, полный охват

→ Разметка переключения кодов (в основном неоднословные фрагменты)

- во всех корпусах, кроме энецкого
- последовательно – для камасинского (А. В. Архипов), в остальных – для части коллекции
- переключение кодов не только с русским

Разметка переключения кодов в проекте INEL



Разметка переключения кодов в проекте INEL

→ Разметка

RUS:ext (utterance-external)

Babuska, tɨ zd’es’?

- предложение или больше

RUS:int.ins (utterance-internal, insertion)

- внутри предложения: границы переключения совпадают с границами составляющих

*Nuŋan bələmn’i **gla:vnɨj g’eroj**.*
‘It is the helper of the protagonist.’

RUS:int.alt (utterance-internal, alternation)

- внутри предложения: границы переключения не совпадают с границами составляющих

Обобщение: идеальный корпус малого языка для исследования переключения кодов

→ С максимальным охватом разных категорий носителей – корпуса вымирающих языков?

→ С максимальным охватом текстов, в т.ч. не исключающий тексты с переключением кодов – корпуса вымирающих языков?

→ С подробными метаданным

→ С полными расшифровками в т.ч. переключенных фрагментов

→ С аккуратным и последовательным делением на предложения

→ С возможностью поиска переключенных фрагментов – необязательно специальная разметка!

→ Если есть разметка переключений кодов – то максимально простая и теоретически нейтральная

Наши работы, упомянутые в докладе

Dyachkov V., Khomchenkova I., Pleshak P., & N. Stoynova. 2020. Annotating and exploring code-switching in four corpora of minority languages of Russia // Computational Linguistics and Intellectual Technologies, 20. Papers from the Annual International Conference “Dialogue” (2020). 2020. P. 228–240. (http://www.dialog-21.ru/media/5085/dyachkovvvplusetal-101.pdf)

Stoynova, Natalia. 2020. Inter-speaker variation in code-switching in the situation of language shift: the case of Nanai and Ulch // CELEA-1, Venice/online, 02-03.09.2020.

Stoynova, Natalia. 2021. Language-inherent variability or contact-induced change? The clitic же ‘after all’ in Russian speech of Nanai and Ulcha speakers. 54th Annual Meeting of the Societas Linguistica Europaea, WS “Integrating sociolinguistics and typological perspectives on language variation” (Athens/online, 30.08-03.09.2021) http://www.sle2021.eu/downloads/SLE%202021%20Extended%20schedule%20final.pdf