Использование полевых записей в задачах обработки речи

Е. Л. Клячко (НИУ ВШЭ, ИЯз РАН), О. А. Сериков (НИУ ВШЭ, DeepPavlov, Институт ИИ)

КОНФЕРЕНЦИЯ «ПОЛЕВЫЕ ЗАПИСИ ЗВУКА: МУЗЫКА, РЕЧЬ, ЛАНДШАФТ» 07.10.2021

РНФ 17-18-01649

Запись озвученного словаря с К. Д. Хутокогиром (Тутончаны, ЭМР, 2008 г.) (http://siberian-lang.srcc.msu.ru)

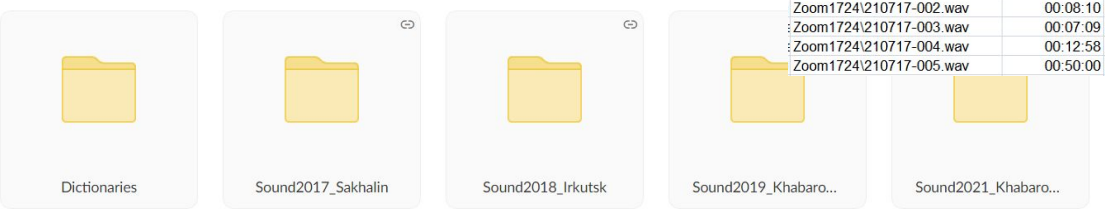
Мотивация

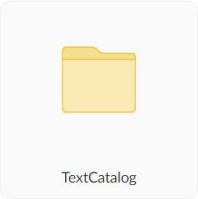
Данные из лингвистических экспедиций

- часто доступны узкому кругу исследователей
- используются с исследовательскими, (реже) педагогическими целями

Возможное использование

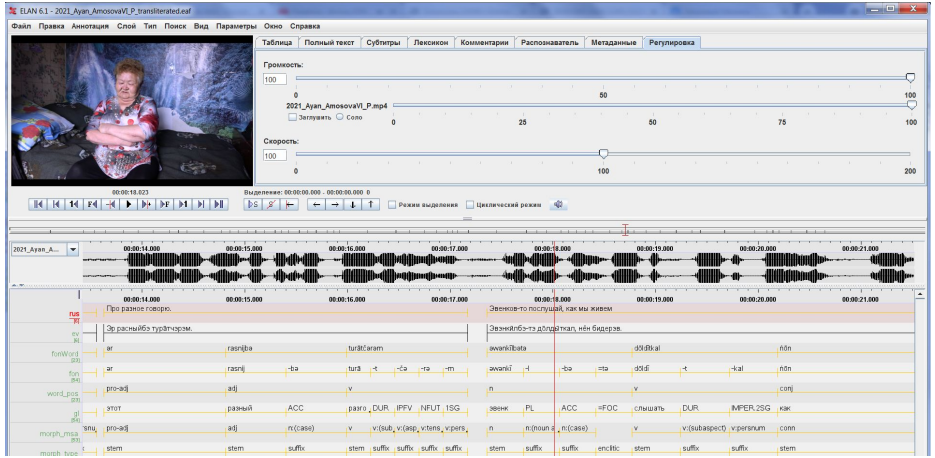
E	F	G	H
имя файла	длительнос	описание	тип материала
TascamDR40Karina\TASCAM_0	00:28:59	охотничий словарь, анкеты, малый словарь [нач	охотничий словарь, анкеты, малый словарь
TascamDR40Karina\TASCAM_0	01:38:55	малый словарь [окончание]	малый словарь
TascamDR40Karina\TASCAM_0	00:55:14	охотничий словарь, анкеты, разговор по-русски	охотничий словарь, анкеты, разговор по-русски
TascamDR40Karina\TASCAM_0	00:09:02	разговор по-русски об огородах, о реках	разговор по-русски
Zoom1725\210716-000.wav	00:41:56	рассказ по-эвенкийски	рассказ по-эвенкийски
DR05\DR0000_0094.wav	00:30:20	рассказ по-эвенкийски	рассказ по-эвенкийски
DR05\DR0000_0095.wav	00:04:59	рассказ по-эвенкийски	рассказ по-эвенкийски
DR05\DR0000_0096.wav	00:02:51	рассказ по-эвенкийски	рассказ по-эвенкийски
DR05\DR0000_0097.wav	00:40:23	рассказ по-эвенкийски	рассказ по-эвенкийски
TascamDR40Karina\TASCAM_0	00:57:14	стрелковый словарь	стрелковый словарь
TascamDR40Karina\TASCAM_0:	00:48:51	рассказ по-эвенкийски	рассказ по-эвенкийски
TascamDR40Karina\TASCAM_0:	01:37:40	охотничий словарь, анкеты, малый словарь [кус	охотничий словарь, анкеты, малый словарь
TascamDR40Karina\TASCAM_0:	00:01:07	малый словарь [куски]	малый словарь
TascamDR40Karina\TASCAM_0:	00:01:54	малый словарь [куски]	малый словарь
TascamDR40Karina\TASCAM_0:	00:00:28	малый словарь [расстроился]	малый словарь
TascamDR40Karina\TASCAM_0:	00:00:29	разговор по-русски	разговор по-русски
Zoom1724\210717-001.wav	00:01:52	малый словарь [человек—голова]	малый словарь
Zoom1724\210717-002.wav	00:08:10	малый словарь [голова—нос]	малый словарь
Zoom1724\210717-003.wav	00:07:09	попытка рассказа по-эвенкийски	рассказ по-эвенкийски
Zoom1724\210717-004.wav	00:12:58	песня по-эвенкийски, попытка рассказа по-эвенк	рассказ по-эвенкийски, песня
Zoom1724\210717-005.wav	00:50:00	разговорник, малый словарь	разговорник, малый словарь





- Обучение на данных:
  - упрощение разметки архивов
  - инструменты для языков

Полевые записи в лингвистике



Тексты

- аудио, видео
- монолог vs диалог vs полилог
- спонтанный → направленный
- фольклор VS автобиография

https://www.youtube.com/watch?v=UUKcKkjys9Q

Полевые записи в лингвистике

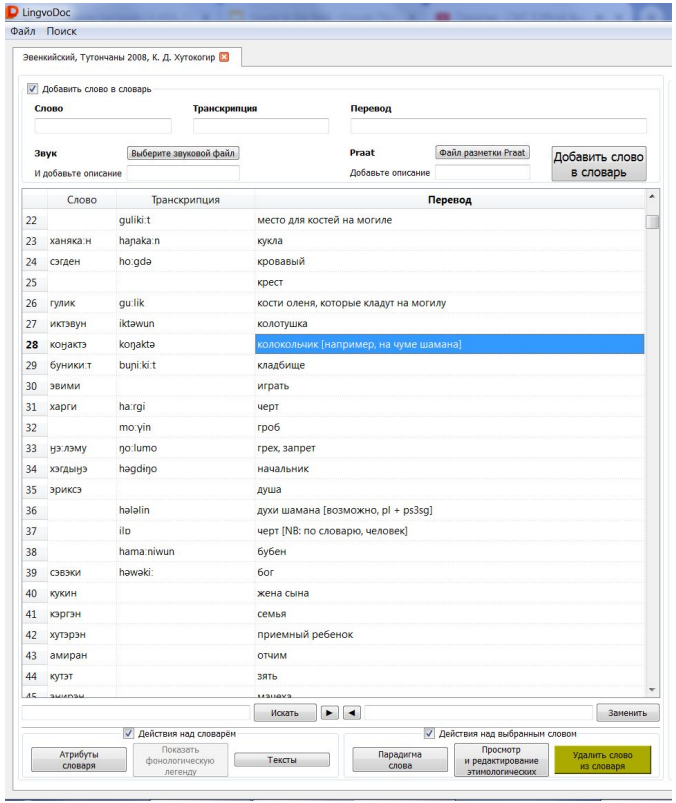
Лексические списки

- аудио
- отдельные слова / словосочетания

Грамматические анкеты

- аудио
- отдельные словосочетания

(перевод, заполнение пропусков)



Полевые записи в лингвистике: хранение

Корпус эвенкийского языка (ЛАЛС НИВЦ МГУ) EN | RU |

Запрос

Слово №1

Слово:

Лемма:

Грамматика:

Глоссы:

Язык/слой: любой

Полнотекстовый поиск: точное соответствие

Поиск предложений | Поиск слов / лексем | Выбор подкорпуса

Результат поиска: найдено

Огонь Ёлдогир Валентина Христофоровна 2007 / URL

albaran ilatčami toyojo .

alba-ra-n ila-t-ča-mi toyo-jo
не.мочь-NFUT-3SG зажечь-DUR-IPFV-INF огонь-ACCIN

Албаран илатчами тогоё.

Не смогла разжечь огонь.

Женщина-шаманка Эспек (Хукочар) Елена Кирилловна 2014 / URL

nuŋanman əfkō hərīwuwrə .

nuŋan-ma-n ə-fkō hərī-wu-w-rə
3SG-ACC-PS3SG NEG-PIMPDEB разбудить-PASS-TR-PNEG

Нуŋанман эвко̄ hэрйвуврэ.

Ее не надо будить

Lingvodoc 3.0

Sign In | Sign Up | Tasks 0 | En

Altaic languages › Manchu-Tungus languages › North Tungus (Tungus proper) › Evenki › Northern Evenki › Evenki, Tutonchana 2008, (K.D. Hutokogir) › Lexical Entries

View published (909) | View contributions (122) | Merge suggestions | Tools | Filter

Word	Phonemic transcription / Narrow phonetic transcription	Meaning	Sound / Markup	Anomalies	Cognates	Comment	Proto-form
илаŋна	ilaŋda		tri_zhivotny... / 126_7325_115...		Cognates		*ilan
ачин		[отрицание при именах]					
су	hu	2pl	2pl.wav... / 126_7296_990...		Cognates		*sū

Существующие коллекции

https://www.paradisec.org.au
https://dobes.mpi.nl

https://elar.soas.ac.uk
http://lingvodoc.ispras.ru

Задачи в обработке звучащей речи

- идентификация языка
- STT
- TTS
- …

Плюсы и минусы полевых коллекций

–

- Нет единообразия в аннотации
- Часто один аннотатор
- Лицензирование?

+

- Хорошая аннотация
- Большая вариативность (диалекты и т. п.)
- Иногда: единственный доступный источник

Две дорожки на наших данных

задача \ дорожка	SigTyp	LowRec2021
транскрибирование	нет	да
количество говорящих	нет	да
наличие разных языков	нет	да
категоризация языков	да	да
категоризация языков по группам	нет	да
категоризация языков по семьям	нет	да

Соревнование Диалог-2021

https://lowresource-lang-eval.github.io/

Полевые данные из Lingvodoc

Ссылка	Язык	Диалект	Группа	Семья	Место	Год записи	Спикер	Спикер, возраст	Количество фрагментов	Категория	Другие ресурсы по языку	Коммента
http://lingvodoc.ispras.ru/dictionary/1108/988/perspective/1108/989/view	карельский	ливвиковский	финно-угорские	уральские	?	?	ж		390?			
http://lingvodoc.ispras.ru/dictionary/1552/1256/perspective/1552/1257/view	финский	Йоэнсуу?	финно-угорские	уральские	?	?	м		100+			
http://lingvodoc.ispras.ru/dictionary/770/7892/perspective/770/7894/view http://lingvodoc.ispras.ru/dictionary/770/7892/perspective/770/7893/view	водский	песоцко-лужицкий	финно-угорские	уральские	?	?	несколько спикеров, но можно различить по именам файлов	несколько спикеров, но можно различить по именам файлов	3000+			

CodaLab

My Competitions Help oserikov

Max submissions per day: 999

Max submissions total: 3

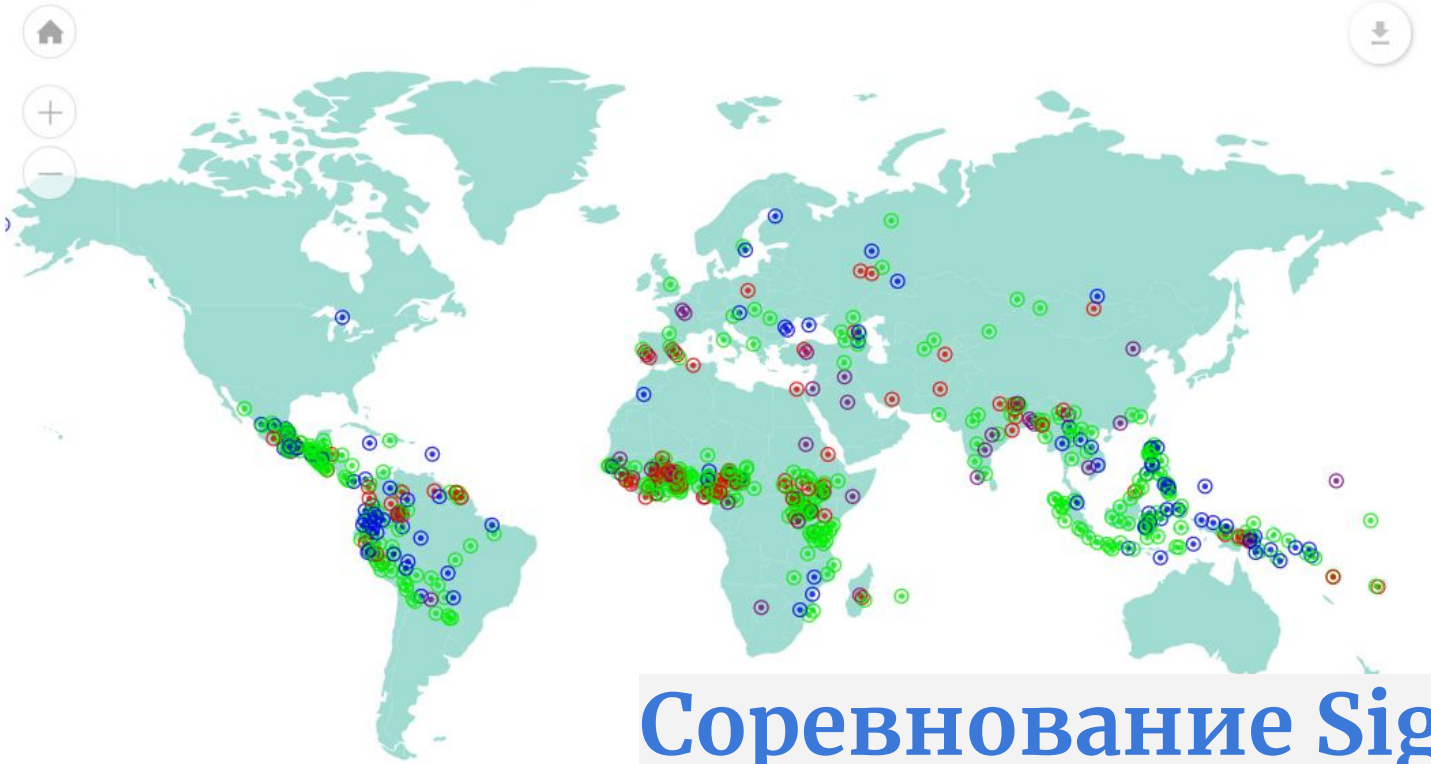
Submissions

Download CSV

#	SUBMITTED	SUBMITTED BY	SUBMISSION ID	FILENAME	STATUS	LEADERBOARD	RESULTS					
1	March 17, 2021, 9 a.m.	gisly	833133	res.zip	Finished	True	0.0125	+	DEL	HIDE	FAILED	RE-RUN
2	March 19, 2021, 9:44 p.m.	NTR	836290	res.zip	Failed	False	---	+	DEL	SHOW	FAILED	RE-RUN
3	March 19, 2021, 9:48 p.m.	NTR	836291	res.zip	Failed	False	---	+	DEL	SHOW	FAILED	RE-RUN
4	March 19, 2021, 9:58 p.m.	NTR	836295	res.zip	Failed	False	---	+	DEL	SHOW	FAILED	RE-RUN
5	March 19, 2021, 10:16 p.m.	NTR	836306	res.zip	Finished	False	0.0462	+	DEL	SHOW	FAILED	RE-RUN
6	March 23, 2021, 3:27 p.m.	dangrebenkin	840915	input_task2.csv.zip	Failed	False	---	+	DEL	SHOW	FAILED	RE-RUN
7	March 23, 2021, 3:38 p.m.	dangrebenkin	840921	res.zip	Failed	False	---	+	DEL	SHOW	FAILED	RE-RUN

CMU Wilderness Multilingual Speech Dataset: MAP

Currently 699 languages completed.
Back to Language List
Dot color indicates alignment quality.
Blue: Very Good; Green: Good; Red: Okay; Purple: Not Okay; Yellow not processed yet



Соревнование SigTyp 2021

Соревнование Диалог-2021

	text	recognized	distance	len
		uraḃalɪ	8	4
		waotɕiuəaotɕiuəai	15	6
2	läʃʃe	læs:opˈesolˈeʂə	15	5
3	ʃuio	ʂuəio	2	4
4	ʃuo	ʂoʂuo	3	3
...	...	...	...	...
8527	ďalitpi	ɒˈlɪpɪpˈɪlˈɛkɔrɒa	16	7
8528	ibďa:ďənnə	ɪnikadˈei:ikætˈi:	15	10
8529	əwilibgɒ:ďəm	ḃilˈɪxpuəɔdˈjɛmɒmˈi:lˈɪpɔaɲʊəḃɪɲi:puərt̪	36	12
8530	su:ďərən	tɒzeʁɒlʊpol̪d͡ʒiʁɪtˈɪd͡ʒ	24	8
8531	ba:sa:məm	mas:ahaɔrasbaɪrimɔzɛt	16	9

- Определить язык + генеалогическую принадлежность по звучащему фрагменту
- Транскрибировать языковые данные

Описание систем команд

NTR/TSU:

-MFCC

-CNN with a self-attentive pooling layer for the classification task (QuartzNet ASR)

-augmentation techniques

Таблица лидеров

Team	LId	GId	FId
NTR	0.06	0.34	0.61
baseline	0.01	0.22	0.82

After the deadline

Team	LId	GId	FId
alumae	0.24	0.63	0.79
NTR	0.06	0.34	0.61
baseline	0.01	0.22	0.82

Automatic language and family detection scores

Team alumae

- architecture: ResNet-50 w. attention + dense classification layer
- pretrained (!) on **VoxLingua107** (arxiv.org/abs/2011.12998)

Выводы

- полевые vs неполевые
  - качество неполевых данных
- классификация языков VS классификация спикеров
- важность первоначальной подготовки данных:
  - проще с самого начала сделать хорошо, чем переразмечать
- важность сырых данных?
- отсутствие лицензий, контактов
- несогласованность интересов
- этические вопросы: как объяснить носителю языка, для чего могут быть использованы его данные

Вопросы?