glottothèque

languages of Anatolia, Caucasus, Iran, Mesopotamia

Georgian

spoken data

Stavros Skopeteas

University of Göttingen

Göttingen, October 25, 2021

visit glottothèque at: https://spw.uni-goettingen.de/projects/lacim/

targets and contents

Goal

available Georgian corpus data

Contents

- resources with written texts
- Georgian spoken data
- recommendations

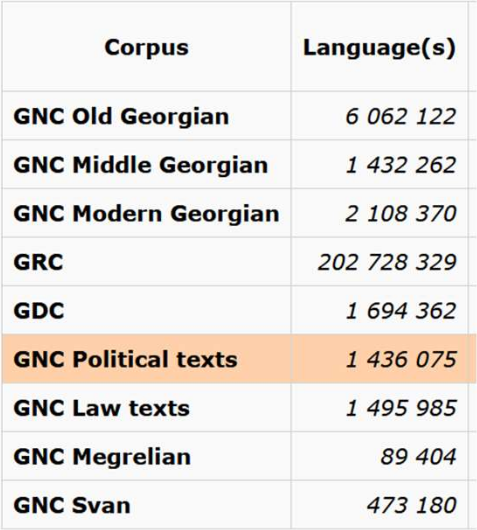
written data

Georgian National Corpus (GNC)

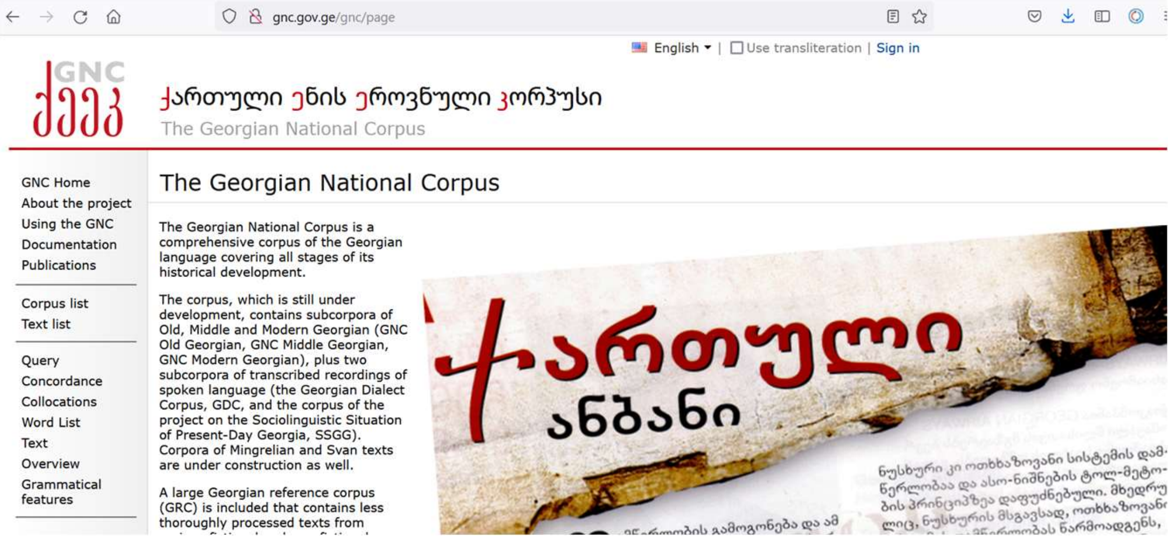
Georgian

rich collection of texts from various diachronic stages of Georgian (Gippert & Tandashvili 2015)

http://gnc.gov.ge/gnc/page



Corpus	Language(s)
GNC Old Georgian	*6 062 122*
GNC Middle Georgian	*1 432 262*
GNC Modern Georgian	*2 108 370*
GRC	*202 728 329*
GDC	*1 694 362*
GNC Political texts	*1 436 075*
GNC Law texts	*1 495 985*
GNC Megrelian	*89 404*
GNC Svan	*473 180*



ქართული ენის ეროვნული კორპუსი

The Georgian National Corpus

The Georgian National Corpus

The Georgian National Corpus is a comprehensive corpus of the Georgian language covering all stages of its historical development.

The corpus, which is still under development, contains subcorpora of Old, Middle and Modern Georgian (GNC Old Georgian, GNC Middle Georgian, GNC Modern Georgian), plus two subcorpora of transcribed recordings of spoken language (the Georgian Dialect Corpus, GDC, and the corpus of the project on the Sociolinguistic Situation of Present-Day Georgia, SSGG). Corpora of Mingrelian and Svan texts are under construction as well.

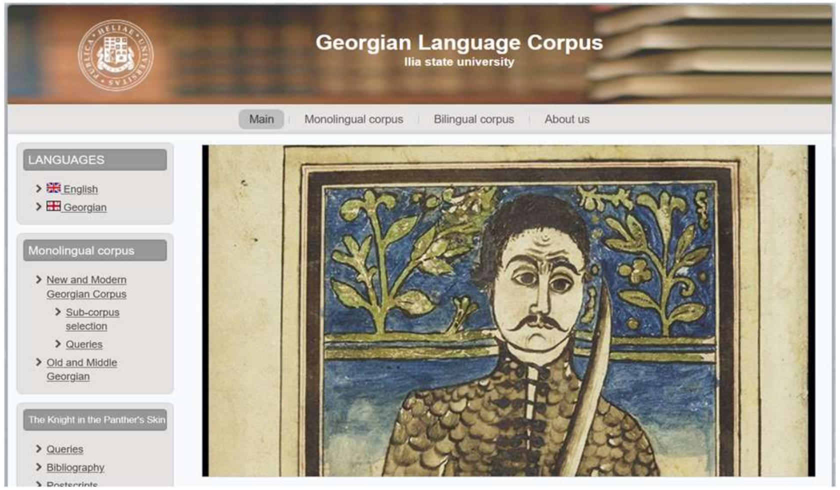
A large Georgian reference corpus (GRC) is included that contains less thoroughly processed texts from

Georgian Language Corpus, ILIA Univ. resources

Georgian Language Corpus

http://corpora.iliauni.edu.ge/

texts of Modern Georgian (1832-2012), Old Georgian section, bilingual section with parallel corpora (Old Georgian – Old Armenian) such as the Georgian Chronicles (Doborjginidze & Lobzhanidze 2016).



Epigraphic Corpus of Georgia

https://epigraphy.iliauni.edu.ge/en/

digitized ancient inscriptions in Greek, Aramaic, Armenian and Georgian (5 c. BC - 19 c. CE; Doborjginidze & Kalkhitashvili 2019)

Wardrops' Collection Online

http://manuscript.iliauni.edu.ge/index.html

manuscript collection with an elaborate structure of textual metadata (Lobzhanidze 2018)

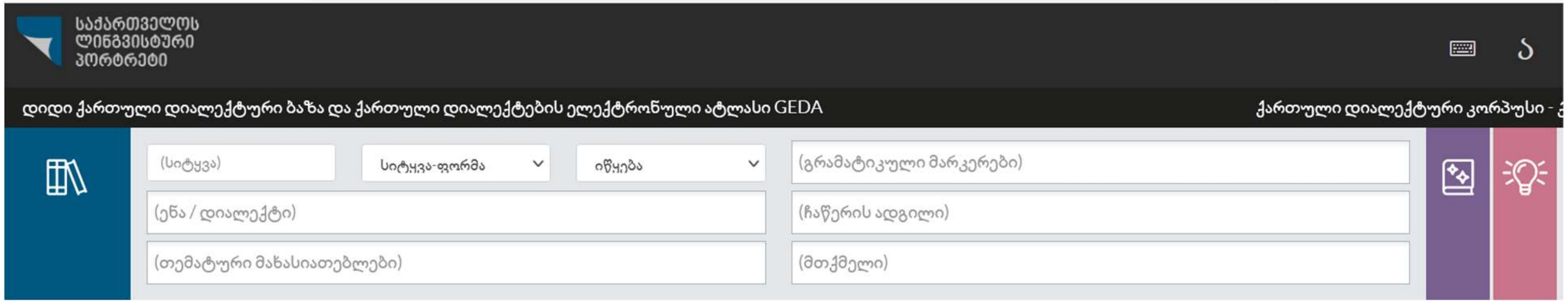
Georgian dialect corpus

http://www.corpora.co/#/corpus

created at the TSU Arnold Chikobava Institute of Linguistics

- almost all dialects of Georgian
- Kartvelian languages

interface and contents in Georgian (also including a database of the Georgian linguistic terms)



ᲡᲐᲥᲐᲠᲗᲕᲔᲚᲝᲡ ᲚᲘᲜᲒᲕᲘᲡᲢᲣᲠᲘ ᲞᲝᲠᲢᲠᲔᲢᲘ

დიდი ქართული დიალექტური ბაზა და ქართული დიალექტების ელექტრონული ატლასი GEDA

ქართული დიალექტური კორპუსი -

სიტყვა-ფორმა

იწყება

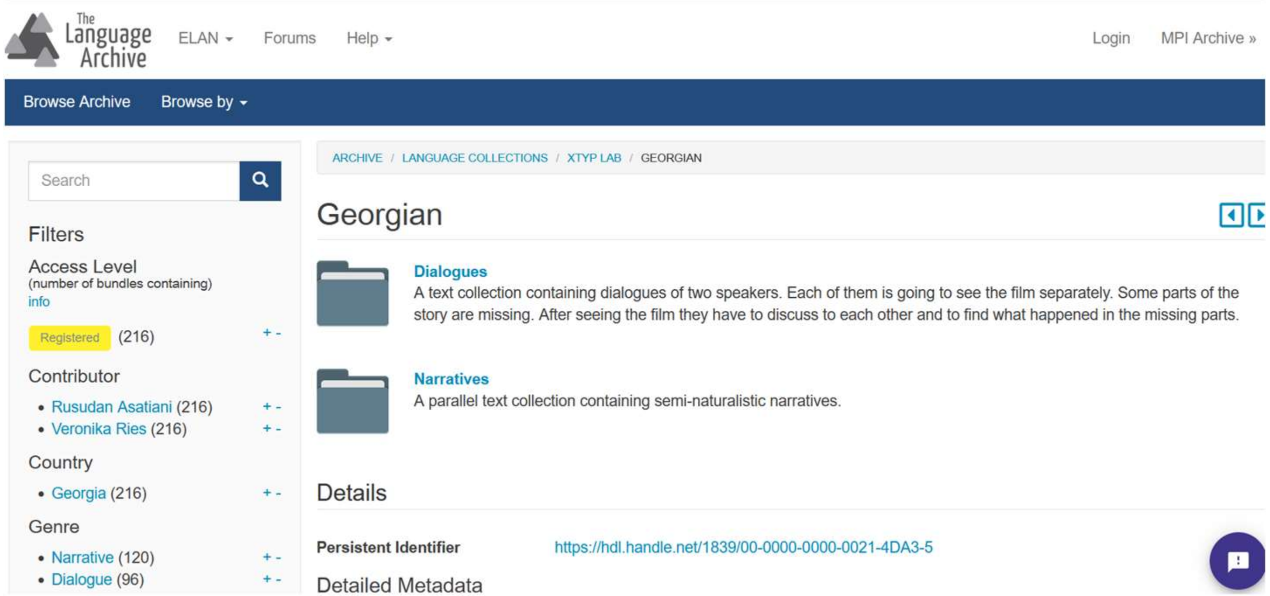
http://www.corpora.co/#/corpus

spoken data

spoken data

Asatiani, Rusudan (recording/transcription,annotation) Stavros Skopeteas (design/supervision), Veronika Ries (recordings), Caroline Brokmann and Florian Fischer (revisions) 2019. Georgian spoken data corpus. *The Language Archive*, Corpus resource;

persistent identifier: https://hdl.handle.net/1839/00-0000-0000-0021-4DA3-5



ARCHIVE / LANGUAGE COLLECTIONS / XTYP LAB / GEORGIAN

Georgian

Dialogues
A text collection containing dialogues of two speakers. Each of them is going to see the film separately. Some parts of the story are missing. After seeing the film they have to discuss to each other and to find what happened in the missing parts.

Narratives
A parallel text collection containing semi-naturalistic narratives.

Filters

Access Level
(number of bundles containing)
info

Registered (216) + -

Contributor
- Rusudan Asatiani (216) + -
- Veronika Ries (216) + -

Country
- Georgia (216) + -

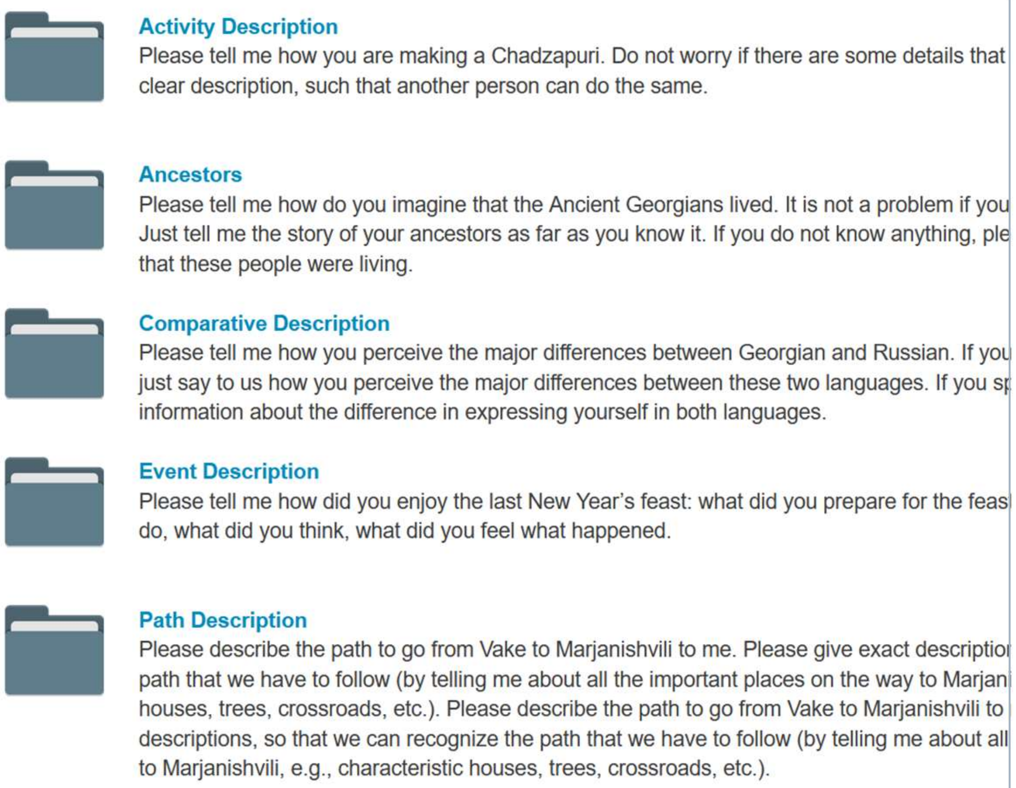
Genre
- Narrative (120) + -
- Dialogue (96) + -

Details

Persistent Identifier https://hdl.handle.net/1839/00-0000-0000-0021-4DA3-5

Detailed Metadata

parallel corpus design



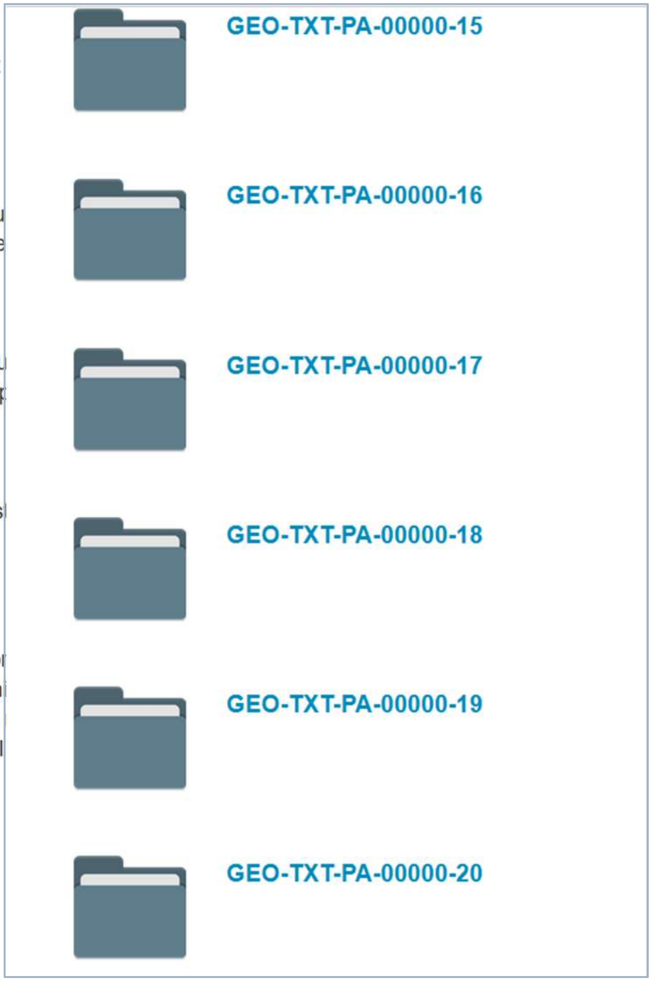
Activity Description
Please tell me how you are making a Chadzapuri. Do not worry if there are some details that
clear description, such that another person can do the same.

Ancestors
Please tell me how do you imagine that the Ancient Georgians lived. It is not a problem if you
Just tell me the story of your ancestors as far as you know it. If you do not know anything, ple
that these people were living.

Comparative Description
Please tell me how you perceive the major differences between Georgian and Russian. If you
just say to us how you perceive the major differences between these two languages. If you sp
information about the difference in expressing yourself in both languages.

Event Description
Please tell me how did you enjoy the last New Year’s feast: what did you prepare for the feas
do, what did you think, what did you feel what happened.

Path Description
Please describe the path to go from Vake to Marjanishvili to me. Please give exact descriptio
path that we have to follow (by telling me about all the important places on the way to Marjan
houses, trees, crossroads, etc.). Please describe the path to go from Vake to Marjanishvili to
descriptions, so that we can recognize the path that we have to follow (by telling me about all
to Marjanishvili, e.g., characteristic houses, trees, crossroads, etc.).



GEO-TXT-PA-00000-15

GEO-TXT-PA-00000-16

GEO-TXT-PA-00000-17

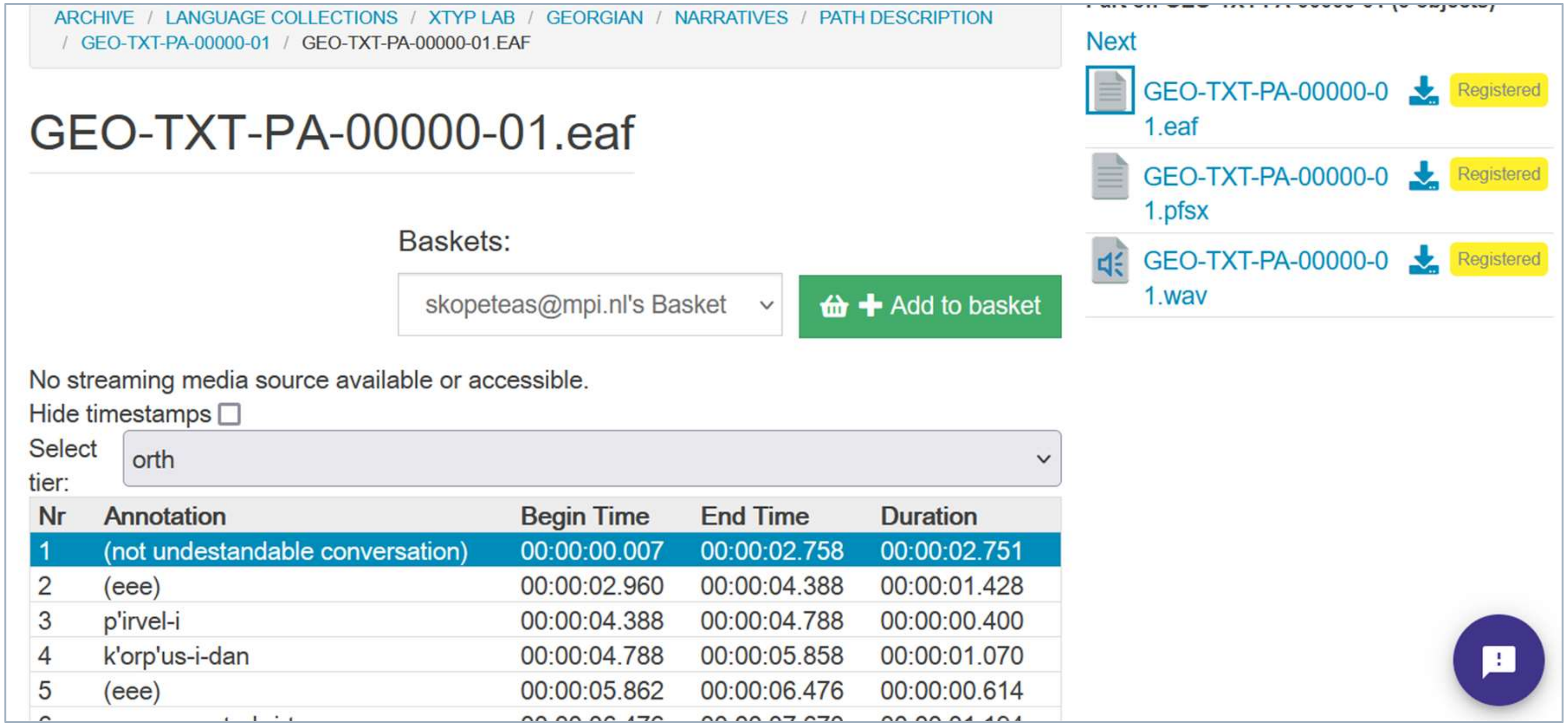
GEO-TXT-PA-00000-18

GEO-TXT-PA-00000-19

GEO-TXT-PA-00000-20

download files

all data are accessible to registered users in the TLA (you need to create an TLA account and login in order to download this data)



ARCHIVE / LANGUAGE COLLECTIONS / XTYP LAB / GEORGIAN / NARRATIVES / PATH DESCRIPTION / GEO-TXT-PA-00000-01 / GEO-TXT-PA-00000-01.EAF

GEO-TXT-PA-00000-01.eaf

Baskets: skopeteas@mpi.nl's Basket

No streaming media source available or accessible.

Hide timestamps □

Select tier: orth

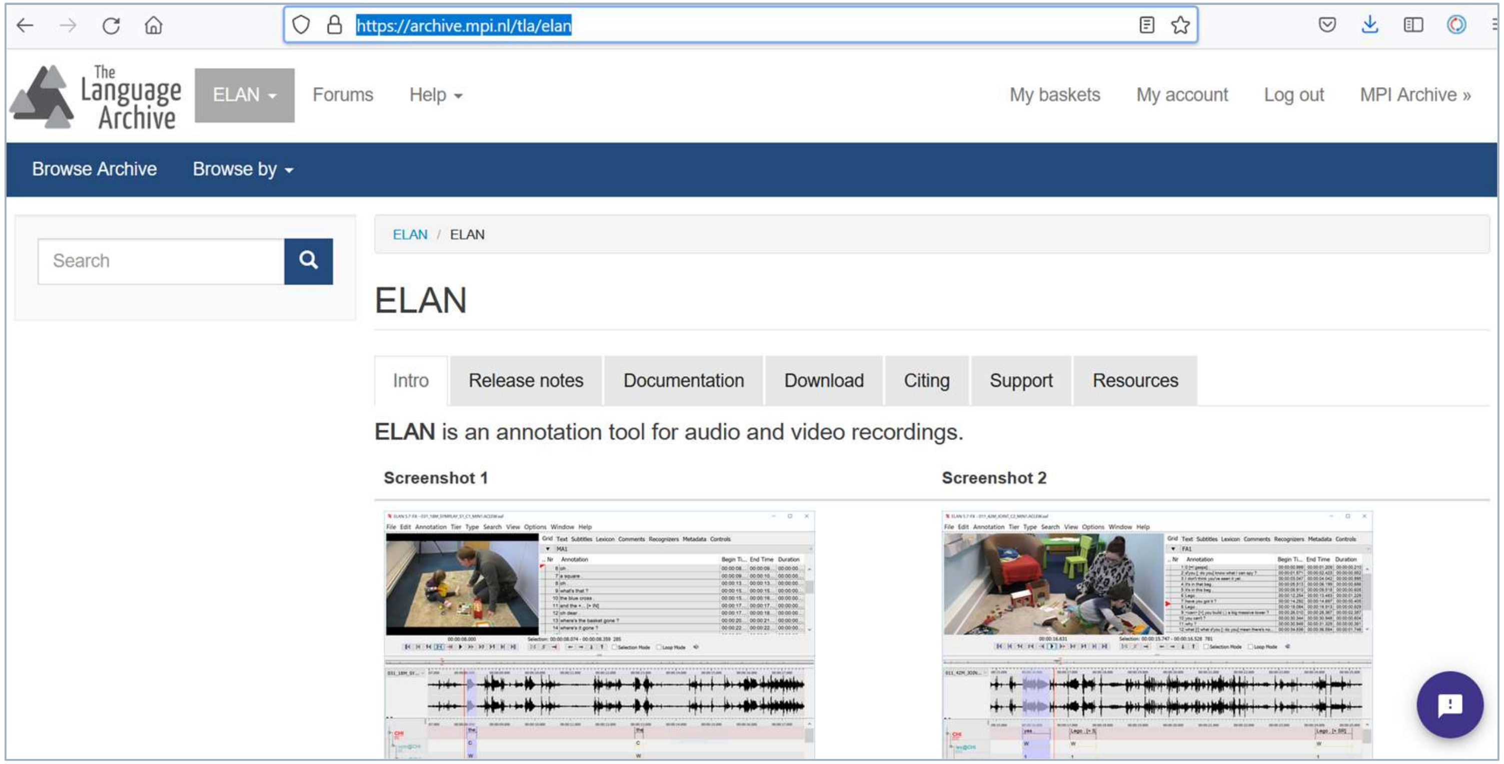
Nr	Annotation	Begin Time	End Time	Duration
1	(not undestandable conversation)	00:00:00.007	00:00:02.758	00:00:02.751
2	(eee)	00:00:02.960	00:00:04.388	00:00:01.428
3	p'irvel-i	00:00:04.388	00:00:04.788	00:00:00.400
4	k'orp'us-i-dan	00:00:04.788	00:00:05.858	00:00:01.070
5	(eee)	00:00:05.862	00:00:06.476	00:00:00.614

GEO-TXT-PA-00000-01.eaf		Registered
GEO-TXT-PA-00000-01.pfsx		Registered
GEO-TXT-PA-00000-01.wav		Registered

open files

download ELAN

https://archive.mpi.nl/tla/elan

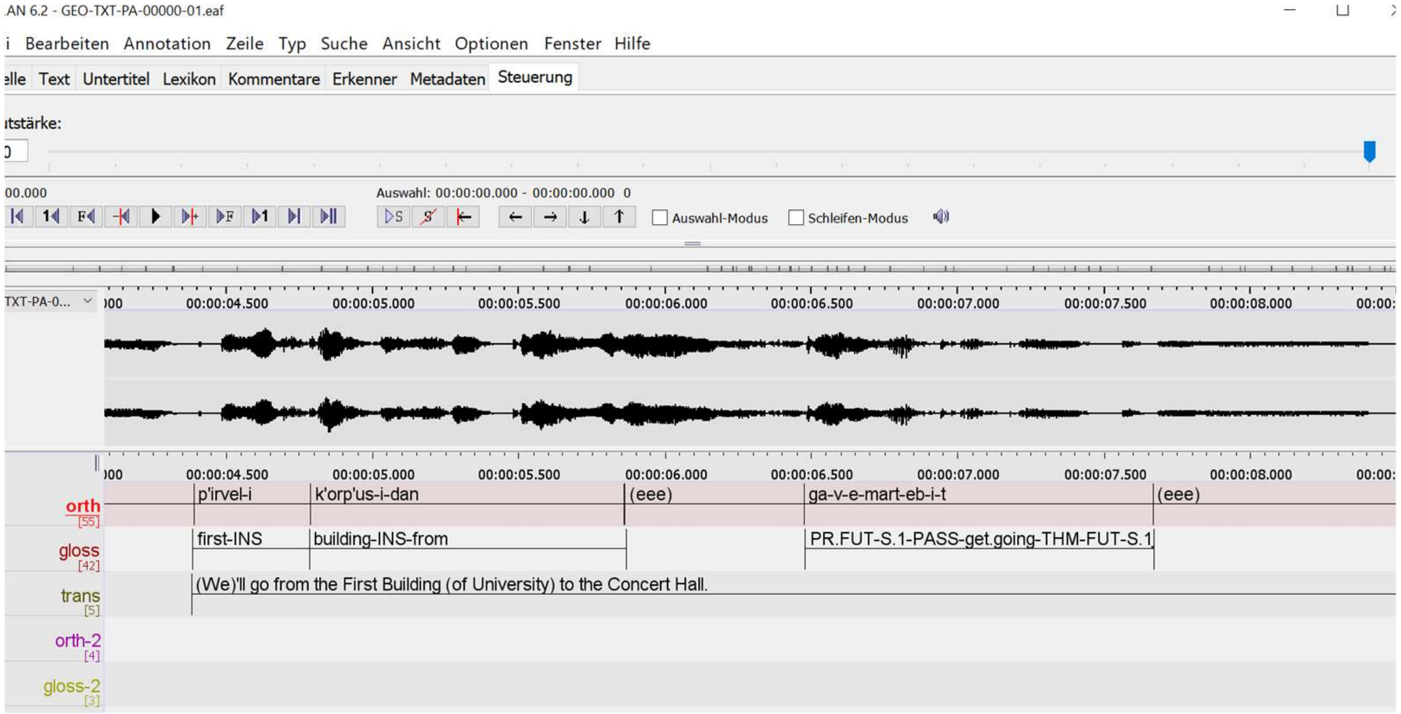


ELAN

ELAN is an annotation tool for audio and video recordings.

process files

- phonological transcription (orth)
- glossing in English
- free translation (English)



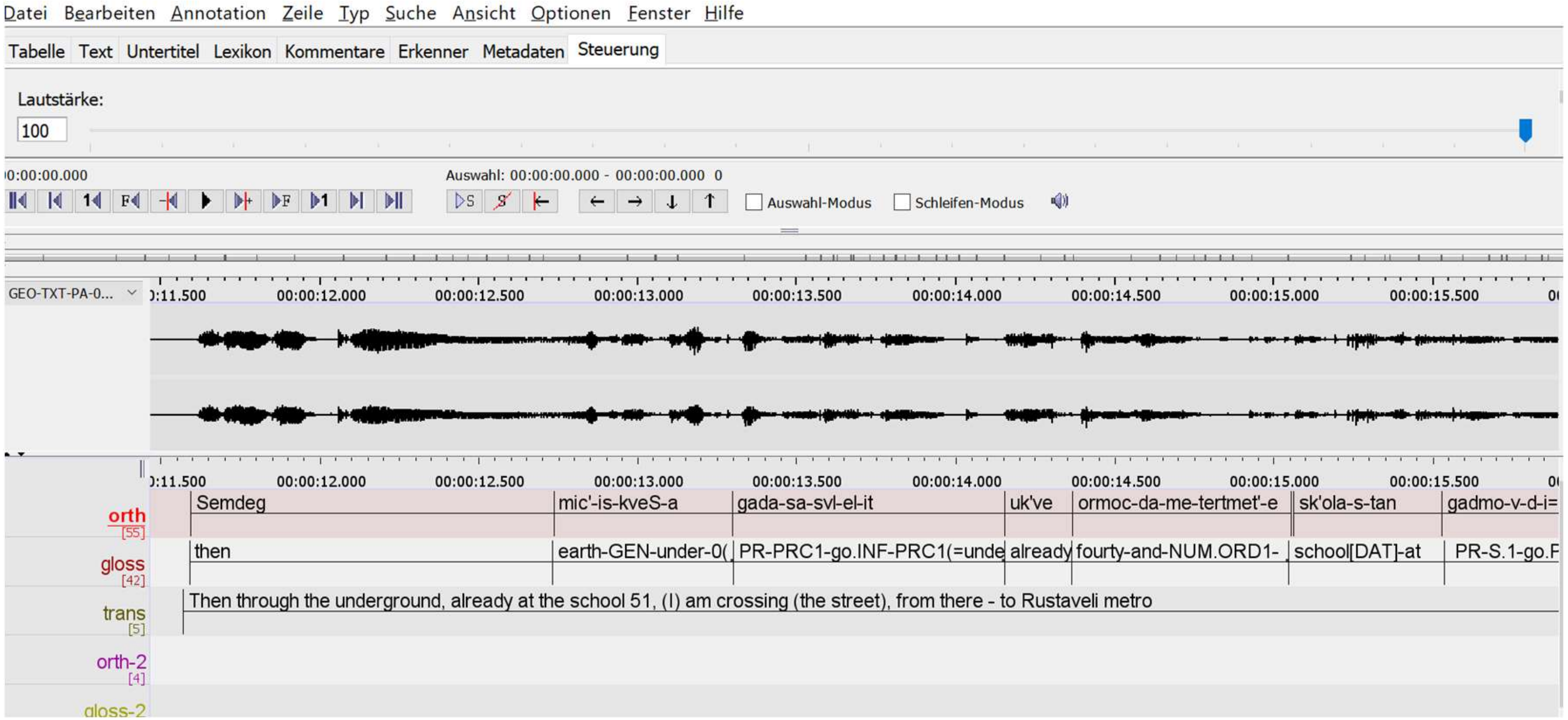
Bearbeiten Annotation Zeile Typ Suche Ansicht Optionen Fenster Hilfe
elle Text Untertitel Lexikon Kommentare Erkenner Metadaten Steuerung
tstärke:
00.000
Auswahl: 00:00:00.000 - 00:00:00.000 0
Auswahl-Modus
Schleifen-Modus
TXT-PA-0...
00:00:04.500
00:00:05.000
00:00:05.500
00:00:06.000
00:00:06.500
00:00:07.000
00:00:07.500
00:00:08.000
orth-2 [4]
gloss-2 [3]

AN 6.2 - GEO-TXT-PA-00000-01.eaf

orth [55]	p'irvel-i	k'orp'us-i-dan	(eee)	ga-v-e-mart-eb-i-t	(eee)
gloss [42]	first-INS	building-INS-from		PR.FUT-S.1-PASS-get.going-THM-FUT-S.1	
trans [5]	(We)'ll go from the First Building (of University) to the Concert Hall.				

process files

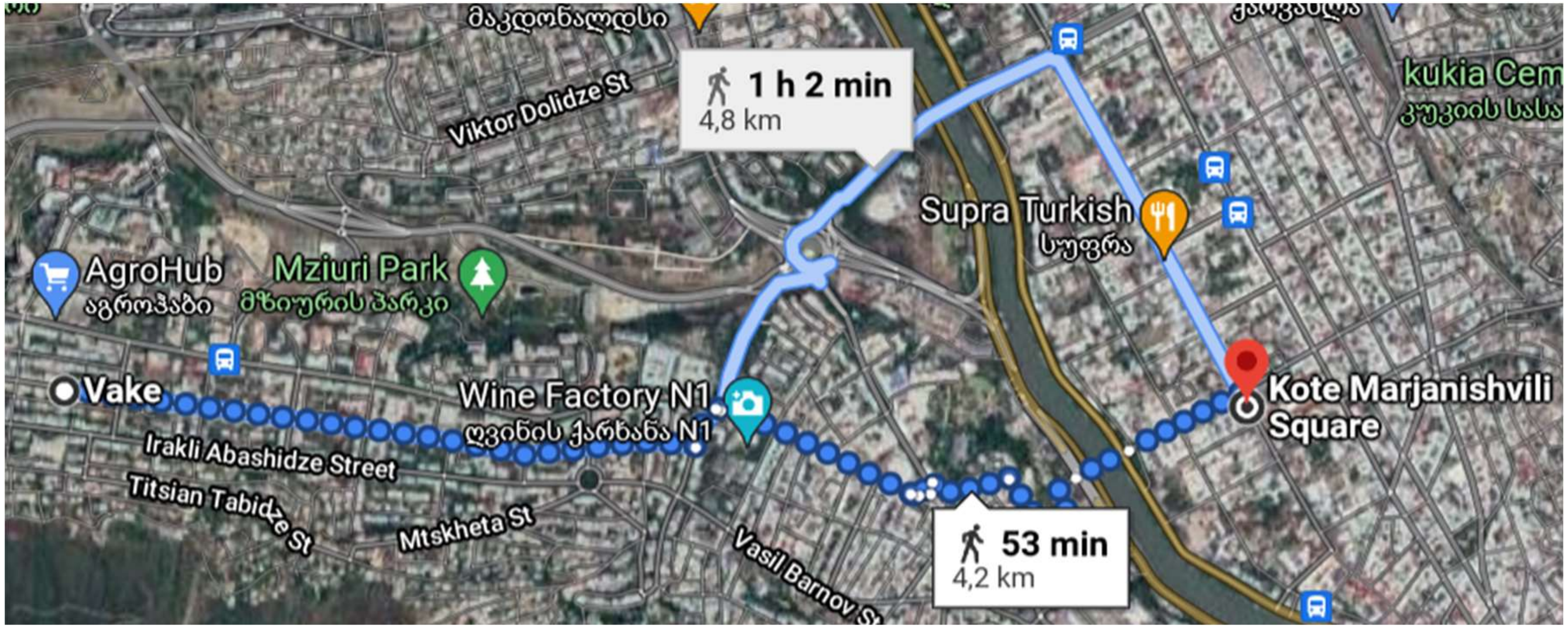
- phonological transcription (orth)
- glossing in English
- free translation (English)


Datei Bearbeiten Annotation Zeile Typ Suche Ansicht Optionen Fenster Hilfe
Tabelle Text Untertitel Lexikon Kommentare Erkenner Metadaten Steuerung
Lautstärke:
100
)0:00:00.000
Auswahl: 00:00:00.000 - 00:00:00.000 0
□ Auswahl-Modus
□ Schleifen-Modus
GEO-TXT-PA-0...
):11.500
00:00:12.000
00:00:12.500
00:00:13.000
00:00:13.500
00:00:14.000
00:00:14.500
00:00:15.000
00:00:15.500
0(

orth [55]	Semdeg	mic'-is-kveS-a	gada-sa-svl-el-it	uk've	ormoc-da-me-tertmet'-e	sk'ola-s-tan	gadmo-v-d-i=
gloss [42]	then	earth-GEN-under-0(	PR-PRC1-go.INF-PRC1(=unde	already	fourty-and-NUM.ORD1-	school[DAT]-at	PR-S.1-go.F
trans [5]	Then through the underground, already at the school 51, (I) am crossing (the street), from there - to Rustaveli metro						
orth-2 [4]							
gloss-2							

exercise

- What do people describe on they way from Vake to Marjanishvili Square?



ideas

This corpus is useful for examining phenomena that often occur in discourse, e.g.:

- phonetic realization of segments, syllable structure, intonation
- word order
- grammatical phenomena that frequently appear in this type of data (narratives/short dialogues).

This corpus does not have the necessary coverage for phenomena that are underrespresented in this type of texts, e.g., narratives/dialogues are not rich in commands. Constructions that are relevant for grammatical description and do not often occur in discourse may be poorly represented or not represented at all in a small-size corpus.

references

Asatiani, R., S. Skopeteas, V. Ries, C. Brokmann, F. Fischer 2019. Georgian spoken data corpus. *The Language Archive*, Corpus resource; persistent identifier: https://hdl.handle.net/1839/00-0000-0000-0021-4DA3-5.

Doborjginidze, N. & I. Lobzhanidze. 2016. Corpus of the Georgian Language. *EURALEX 17*, 328-334.

Doborjginidze, N. & T. Kalkhitashvili 2019. The Epigraphic Corpus of Georgia. *Kadmos* 9, 222-233.

Gippert, J. & M. Tandashvili 2015. Structuring a diachronic corpus. The Georgian National Corpus project. In J. Gippert & R. Gehrke (ed.), *Historical Corpora. Challenges and Perspectives.* 305-322. Tübingen: Narr.

Lobzhanidze, I. 2018. Computational Model of the Modern Georgian Language and Search Patterns for an Online Dictionary of Idioms. *TbiLLC 2018*, 187-208.

this lecture

is part of the series *Glottothèque: Languages of the Anatolia, Caucasus, Iran, Mesopotamia; grammatical snippets online*, ed. by. C. Bulut, A. Donabédian-Demopoulos, G. Haig, G. Khan, P. Samvelian, S. Skopeteas, N. Sumbatova. Bamberg/Cambridge/Göttingen/Moskow/Nicosia/Paris: LACIM network.

You may find related lectures and further information at the Glottothèque website at:
https://spw.uni-goettingen.de/projects/lacim/

Anatolia, Caucasus, Iran, Mesopotamia
grammatical snippets online