Università
della
Svizzera
italiana

Faculty
of
Communication,
Culture
and Society

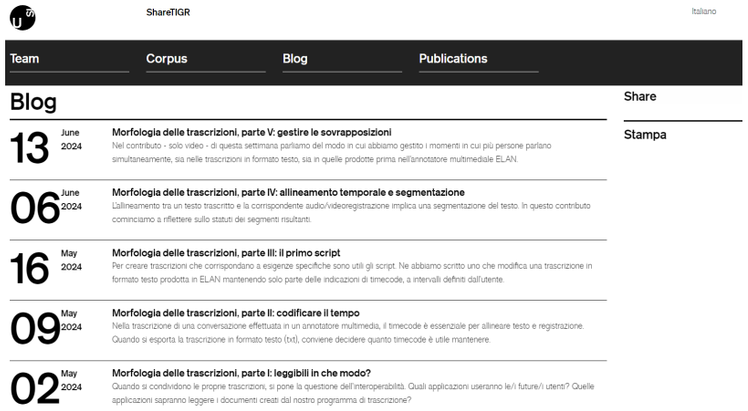
# ShareTIGR - Sharing the TIGR corpus of spoken Italian: an ORD case study

Duration:
February 2024 – January 2025

Funding:
USI Università della Svizzera italiana

Team:
Johanna Miecznikowski (USI)
Elena Battaglia (USI)
Christian Geddo (USI)
Nina Profazi (USI)

# Project goals

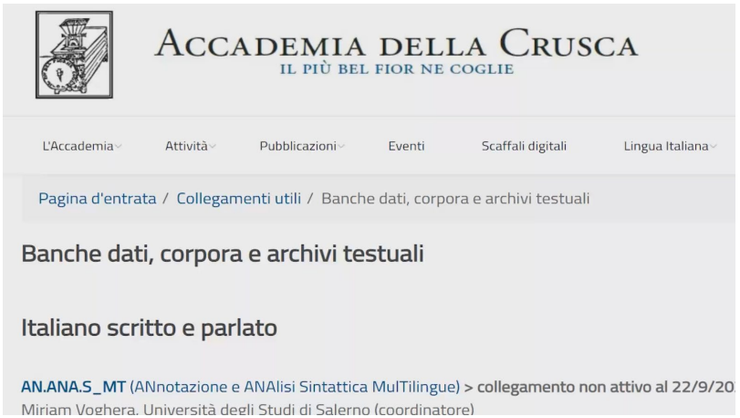




making the TIGR corpus of spoken Italian available for scientific use on LaRS @ SWISSUbase, respecting data protection and FAIR principles (Wilkinson et al. 2016)

(Accessibility via a corpus platform?)

discussing the various phases of this process as **a case study** of open research data practices in linguistics, engaging with potentially interested communities via a lab blog

# The TIGR corpus


Context, design and composition of TIGR

- gathered within InfinIta (SNF grant no. 192771, USI, ‘20-’24) to study evidential means in Italian
- Ticino, Grigioni Italiano
- meal preparation, table conversations, tutoring/ practical classes/lessons, interviews
- Video, audio, GAT 2 transcripts (Selting et al. 2011)
- 23.5 hours, 23 events, 115 speakers

Existing corpora of spoken Italian

- varieties spoken in Italy
- various genres of naturally occurring conversation as well as experimental settings
- transcripts and in some cases audio recordings
- most recent corpus: KiParla (Mauri et al. 2019, kiparla.it)

# Preparing the TIGR corpus to be deposited on the repository

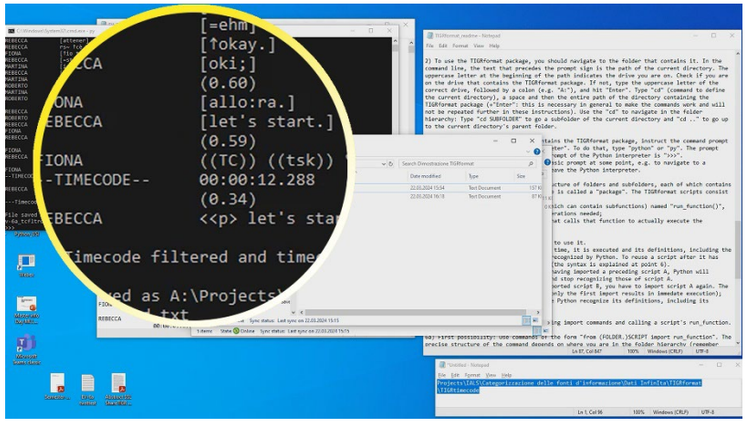

Data processing

- early phase of InfinIta: informed consent
- de-identification of audio (replace names by noise)
- de-identification of video where desired by participants
- Single multimedia files + a A/V files with split-screen image and mixed audio
- script-assisted and manual processing of .txt transcripts for qualitative analysis
- machine-readable transcripts in a format to be defined

Preparation of metadata

Corpus description (website, documents)

## ShareTIGR lab blog




Sharing video data via a repository is not a wide-spread practice in linguistics and raises various challenges when it comes to meet the requirements of open science and at the same time protect personal data (Diaz 2022, Miecznikowski & Profazi 2023).

We reflect on such challenges as they emerge. A lab blog develops these reflections, resulting in an ORD case study in a narrative format. The blog addresses linguists, social scientists, specialists of digital humanities and data management, science writers and the interested public.

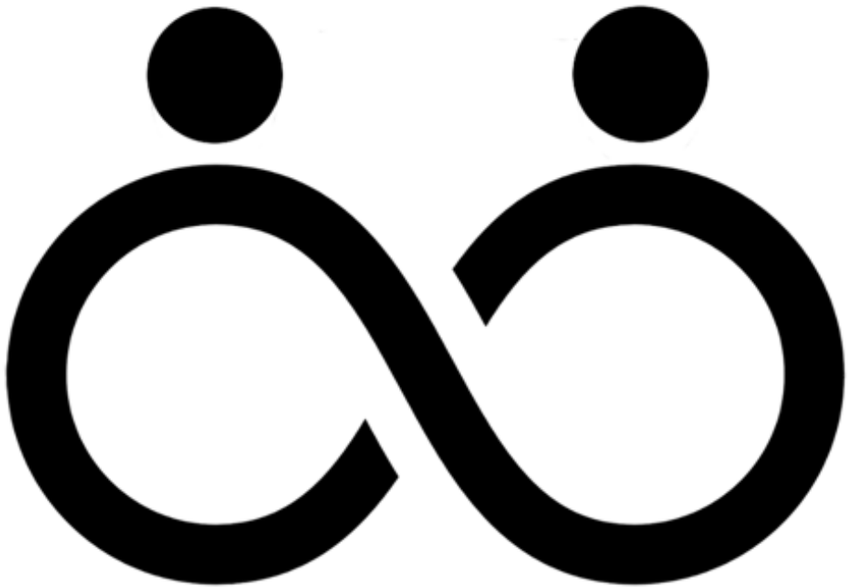
Some blog posts are accompanied by short videos, exploring audiovisual techniques commonly used in video tutorials and essays (cf. the *Progetto culturale* of the USI Facolty of CCS).

## References


- Diaz, P. (2022). Data protection: legal considerations for research in Switzerland. *FORS Guide* No. 17, 1.0. Lausanne: Swiss Centre of Expertise in the Social Sciences FORS. doi.org/10.24449/FG-2022-00017
- Mauri, C., Ballarè, S., Goria, E., Cerruti, M., & Suriano, F. (2019). KIParla corpus: A new resource for spoken Italian. In R. Bernardi, R. Navigli & G. Semeraro (eds.), *Proceedings of the 6th Italian Conference on Computational Linguistics CLiC-it*, CEUR-WS.
- Miecznikowski, J. & Profazi, N. (2023). Spoken language corpora as open research data: the example of KIParla chord-talk-in-interaction.usi.ch/documentation
- Selting, M. et al. (2011). A system for transcribing talk-in-inter-action: GAT 2 translated and adapted for English by Elizabeth Couper-Kuhlen and Dagmar Barth-Weingarten. *Gesprächs-forschung*, 12, 1-51.
- Wilkinson, M. D. et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1), 1-9. doi.org/10.1038/sdata.2016.18


## Websites


- *Accademia della Crusca.* https://accademiadellacrusca.it/it/contenuti/banche-dati-corpora-e-archivi-testuali/6228
- *Corpus KIParla.* kiparla.it/
- *Language Repository of Switzerland LaRS*. lars.uzh.ch/en.html
- *Linguistic Corpus Platform LCP*. lcp.linguistik.uzh.ch/
- *Progetto culturale della Facoltà di CCS*. com.usi.ch/it/progettoculturale.



# InfinIta / ShareTIGR

https://sharetigr.usi.ch/
https://twitter.com/ItaInfin