Università della Svizzera italiana

Faculty of Communication, Culture and society / Istituto di studi italiani

Johanna Miecznikowski (johanna.miecznikowskifuenfschilling@usi.ch), Elena Battaglia (elena.battaglia@usi.ch), Christian Geddo (christian.geddo@usi.ch), Nina Profazi (nina.profazi@usi.ch), corpustigr@usi.ch

13e Journée de Linguistique Suisse, Neuchâtel, 28 juin 2024

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

Overview

ShareTIGR is a one-year project (February 1st, 2024 – January 31st, 2025) funded by USI, which aims at (a) making the TIGR corpus of spoken Italian available for scientific use, respecting data protection and FAIR principles (Wilkinson et al. 2016); (b) discussing the various phases of this process as **a case study** of open research data practices in linguistics, engaging with potentially interested communities via scientific presentations and publications and via a lab blog and social media.

Existing corpora of spoken Italian

To study spoken varieties of Italian, several corpora have been gathered from the 1990s onwards and have partially been made available on websites and DVDs (cf. the corpora listed on Parlaritaliano, Mauri et al. 2019 and the currently growing resources available at https://kiparla.it/). The existing corpora consist of audio recordings and transcripts. Although video-recording has become more common in linguistics (see Mondada 2013) and some video-based research has been carried out recently on Italian talk-in-interaction (e.g. Calabria & De Stefani 2020), no Italian video data are currently widely accessible for research purposes. A further remarkable gap is the absence of accessible corpora for regional Italian varieties spoken in Switzerland. An oral corpus was gathered by the Osservatorio Linguistico della Svizzera italiana in the early 2000s and was used to extract lexical information (Pandolfi 2009), but these data are not accessible to scholars outside OLSI.

Context, design and composition of TIGR

The TIGR corpus consists of 23.5 hours of interactions recorded in Ticino (18 events) and in the Grisons (5 events) in 2021 and 2022. It was gathered within the InfinIta project (SNF grant no. 192771, USI, 2020-2024) to study evidential means in spoken Italian. taking into account various participation frameworks, various types of relation between talk and simultaneous non-verbal activities, and the role of information source in narrative and argumentative discourse. The project goals influenced the corpus design, which includes meal preparation (3-4 participants), table conversations (3-4 participants), tutoring / practical classes/lessons (teacher + 1-20 students), and interviews about the Covid-19 pandemic (2 participants). 115 speakers participated in total, among which 69 residing in Switzerland, 43 in Italy and 3 in other countries. The age range that is best represented is the one between 20 and 29 years and about 3/4 of the speakers finished a higher secondary school or graduated from university.

Despite its origin within the context of a specific investigation, the corpus is suitable for diverse research purposes in linguistics and in the neighbouring disciplines. Its sociolinguistic, diatopic and technical properties define a unique profile in the panorama of corpora for spoken Italian.

InfinIta / ShareTIGR

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

Audio and video recordings

The recordings were made from two different camera angles, up to four clip-on microphones, and 1-2 room microphones, depending on the number of participants. They were synchronized based on timecode using Tentacle Sync audio recorders and external timecode generators for video cameras. This assured a high degree of precision in synchronizing audio and video files.

Transcripts

The recordings were transcribed in ELAN adopting the GAT 2 conventions for fine transcription (Selting et al. 2011), with some adaptations. The transcripts are currently being revised. One deviation from GAT 2 is that, even if prosody is represented, no prosodic segmentation has been carried out. The segmentation in ELAN aimed at facilitating the workflow of transcript revision and of extracting "Jefferson style" transcripts in TXT format from ELAN. In the case of overlap between several speakers, segment boundaries were therefore placed in such a way as to correspond to the square brackets that signal boundaries of overlapping speech in the transcribed text.

Preparing the TIGR corpus for sharing

ShareTIGR aims at turning the corpus into a resource that is reusable in further research. A precondition for this, namely the study participants' informed consent, is fulfilled thanks to the original project's fieldwork protocol. The tasks that remain to be realized in ShareTIGR include further data processing:

- de-identification of audio: replacement of names by noise, voice distortion where desired by participants;
- de-identification of video data where desired by participants (Gaussian blur, 'find edges' effect);
- the production of a single A/V file per event with split-screen image and mixed audio as an additional option besides the single recordings;
- manual and script-assisted processing of transcripts to produce TXT files that are easily readable to the human eye and to certain annotation programs: filtering timecode indications to leave one TC stamp roughly every 10 seconds, checking for orphan square brackets, layout;
- XML transcripts in a format that remains to be defined (TEI standard for spoken language?), which should be interpretable for a wide range of corpus-linguistic applications.

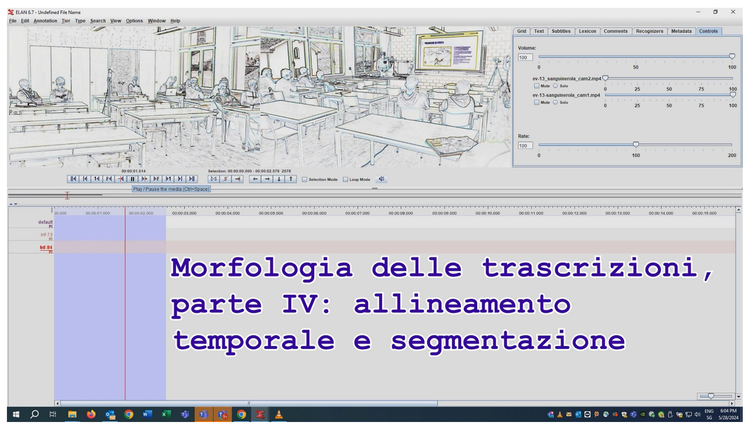
A second set of tasks regards the description of the corpus and the preparation of metadata.

Accessibility via a respository

The data will be deposited on the Language Repository of Switzerland (LaRS), a section of the SWISSUbase repository that offers a metadata scheme specifically designed for linguistic data.

Accessibility via a corpus platform?

The so prepared data will be ready to be made available on a corpus platform as well, i.e. in a server-based digital environment that offers the possibility of viewing, searching, and analysing data online. Platforms for multimedia conversational data are, however, still to be developed in Switzerland. Within a one-year project co-funded by swissuniversities (FAIR-FI-LD, July 1st, 2024-June 30th, 2025, leading house: LiRI / UZH, further partners: Clarin-CH, ZHAW), members of the Share-TIGR team and of USI E-Lab will help adapting LiRI's Linguistic Corpus Platform LCP and the video annotation software Videoscope to create a software prototype for conversational data that can be run on local servers. The TIGR corpus is a possible use case.



ShareTIGR lab blog

Sharing video data via a repository is not a widespread 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 adequately (Diaz 2022, Miecznikowski & Profazi 2023). In ShareTIGR, the team reflects on such challenges as they emerge, using a lab blog in Italian and English and social media. It develops these reflections as an ORD case study in a narrative format to share them with 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).

Corpus web site

A corpus web site is being built, which describes the corpus succinctly in thematic sections and integrates references to the lab blog.

References

- Calabria, V. & De Stefani, E. (2020). Per una grammatica situata: aspetti temporali e multimodali dell'incrementazione sintattica. *Studi Italiani di Linguistica Teorica e Applicata* 49(3), 571-601.
- 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
- Mondada, L. (2013). Video as a tool in the social sciences. In *Handbücher zur Sprach- und Kommunikationswissenschaft* (HSK) 38/1 (pp. 982-992). Berlin, Boston: De Gruyter. doi.org/10.1515/9783110261318.982
- Pandolfi, E. M. (2009). *LIPSI. Lessico di frequenza dell'italiano parlato nella Svizzera italiana*. Bellinzona: Osservatorio linguistico della Svizzera italiana.
- Selting, M. et al. (2011). A system for transcribing talk-in-interaction: GAT 2 translated and adapted for English by Elizabeth Couper-Kuhlen and Dagmar Barth-Weingarten. *Gesprächsforschung*, 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

- *Corpus KIParla.* kiparla.it/
- Language Repository of Switzerland LaRS. lars.uzh.ch/en.html
- Linguistic Corpus Platform LCP. lcp.linguistik.uzh.ch/
- *Parlaritaliano*. paritaliano.studiumdipsum.it/
- Progetto culturale della Facoltà di CCS. com.usi.ch/it/progettoculturale.