

# The TIGR corpus: collection, transcription, processing and management of videorecorded data for the study of Italian talk-in-interaction

Johanna Miecznikowski, Elena Battaglia  
(USI Università della Svizzera italiana)



## Indice

1. Introduzione
2. Perché un corpus video dell’Italiano parlato?
3. Composizione del corpus
4. Parlanti del TIGR
5. Luoghi dove sono stati registrati gli eventi del TIGR
6. Impostazione tecnica delle registrazioni
7. Post-produzione dei dati audio e video
8. Trascrizione allineata ai documenti multimedia
9. La derivazione di trascrizioni in formato testo in stile ‘dialogo teatrale’
10. Verso una versione tokenizzata delle trascrizioni
11. Condivisione dei dati

# 1. Introduzione



**InfinIta** (SNSF grant no. 192771)  
September 2020 – August 2025

**ShareTIGR** (USI ORD program)  
February 2024 – February 2025

Johanna Miecznikowski (PI)  
Elena Battaglia, Christian Geddo (Phd candidates)  
Chiara Sbordoni (research intern)  
Nina Profazi (scientific collaborator)  
Costanza Lucchini, Benedetta Scotto di Santolo, Simona Kaufmann, Tommaso Barenco (student assistants)

## Data to examine the categorisation of information source in talk

Dendale 1994,  
Aikhenvald 2004...

InfinIta examines the categorisation of information source in spoken Italian:

- By which resources (evidential markers and constructions, discourse strategies, multimodal strategies, implicatures) is information source communicated?
- Which distinctions related to information source become relevant?  
Specific source categorisations, e.g. ‘The president is not in good shape. **His speaker said so during yesterday’s press conference / That’s evident from how he behaved during yesterday’s TV debate**’  
vs.  
generic categorisations, e.g. ‘**I heard / Apparently** the president is not in good shape?’
- How does the categorisation of information source contribute to epistemic positioning?

The first goal of TIGR has been to answer those research questions:

- Resources, participant roles and the sequentiality of epistemic positioning vary across **interaction types**. TIGR allows to study that variation.
- TIGR is a **video** corpus and therefore allows to examine multimodal evidential resources.



## 2. Perché un corpus video dell’Italiano parlato?

- Studio dell'uso integrato di mezzi verbali, mezzi multimodali e risorse situazionali
  - fonti in situ (acquisizione di informazioni durante l'interazione in corso grazie alla percezione diretta e alle infeerenze basate su indizi percettivi)
  - ruolo della direzione dello sguardo e di certi gesti nel posizionamento epistemico
  - costruzione del riferimento deittico
  - ...

- Studio dell'uso integrato di mezzi verbali, mezzi multimodali e risorse situazionali
  - fonti in situ (acquisizione di informazioni durante l'interazione in corso grazie alla percezione diretta e alle infeerenze basate su indizi percettivi)
  - ruolo della direzione dello sguardo e di certi gesti nel posizionamento epistemico
  - costruzione del riferimento deittico
  - ...
- Studio della grammatica (sintassi, segnali discorsivi) in interazione:
  - visione più completa dello sviluppo sequenziale, tenendo conto della multi-attività

- Studio dell'uso integrato di mezzi verbali, mezzi multimodali e risorse situazionali
  - fonti in situ (acquisizione di informazioni durante l'interazione in corso grazie alla percezione diretta e alle infeerenze basate su indizi percettivi)
  - ruolo della direzione dello sguardo e di certi gesti nel posizionamento epistemico
  - costruzione del riferimento deittico
  - ...
- Studio della grammatica (sintassi, segnali discorsivi) in interazione:
  - visione più completa dello sviluppo sequenziale, tenendo conto della multi-attività
- Studio della variazione diafasica
  - visione più adeguata della varietà generica (tipi di situazione)

# 3. Composizione del corpus

## The TIGR corpus in and beyond Infinita

TIGR was collected in view of a specific research goal (*special corpus*).

The first goal of TIGR has been to answer research questions related to information source in interaction:

- Resources, participant roles and the sequentiality of epistemic positioning vary across **interaction types**. TIGR allows to study that variation.
- TIGR is a **video** corpus and therefore allows to examine multimodal evidential resources.

However, being varied as to interaction types and being complementary to other corpora of spoken Italian as to geographical variation and technical set-up, it has a potential to be reused in further research.

Teubert &  
Čermáková  
2004:119-120

*Reference corpus* (balanced, representative)  
vs. *opportunistic corpus* (consisting of several *special corpora* “one [sc. the corpus owner/compiler] can lay hands upon”)

## Design del corpus TIGR nel 2020

Tipo di interazione	Length	Events	Total length	Participants	Participants moving	Handling of objects
Setting						
Interview	30'	10	5 h	2-3	-	-
Table conversation	60'	5	5 h	2-4	+	+
Preparing a meal	45'	4	3 h	2-4	++	+
Preparing a presentation	45'	10	7 h 30'	2-4	+	+
Lessons	45'	10	7 h 30'	Teacher and class	-	+

## Il corpus TIGR oggi

- Raccolta dati da maggio 2021 a maggio 2022
- Riadattamenti del design dovuti alle restrizioni Covid e alla disponibilità effettiva dei parlanti

↓  
**23 h 30' di videoregistrazione**

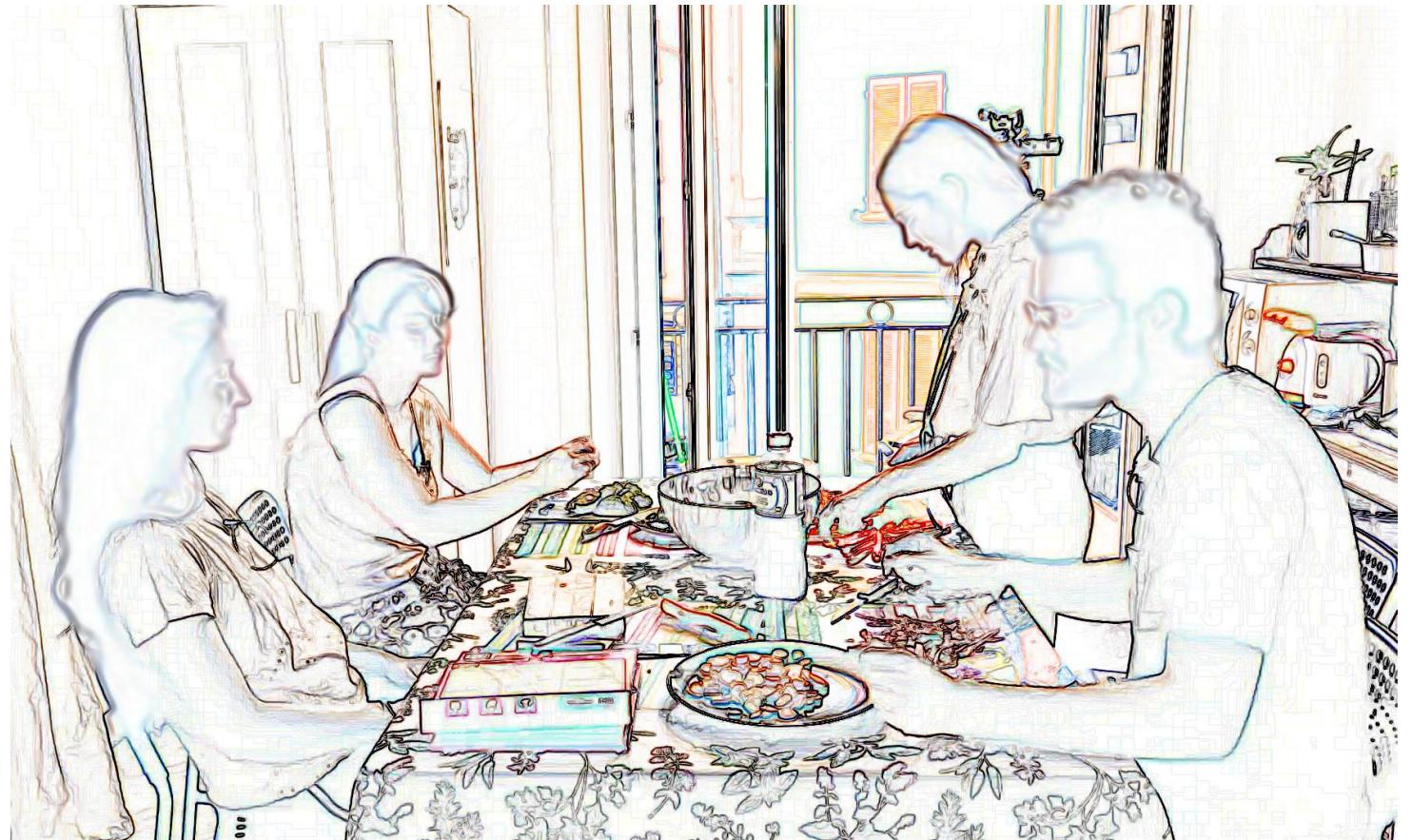
**23 eventi**

**115 partecipanti**

	Table con- versation	Cooking	Educational contexts: lessons and practical instruction			Inter- views
			Tutoring in architecture	Lessons in theater, music, restoration	Lessons in general education and language teaching	
Event-related parameters (Deppermann / Hartung 2011:423-424)						
Number of participants	3-4	4	2-3	6-11	8-19	2
Institutionality	-	-	+	+	+	+
Participant roles	variable	variable	asymmetric	asymmetric	asymmetric	asymm.
Participants moving	+/-	+	+/-	+	+/-	-
Handling of objects	+	+	+	+	+/-	-
Multi-activity	+	++	+	++	-	-
Composition of the corpus						
Events	5	2	4	3	3	6
Length	6 h 05'	1 h 43'	4 h 41'	4 h 51'	2 h 27'	3 h 42'



<b>Conversazioni a tavola</b>	
Parametri	
Numero di partecipanti	3-4
Istituzionalità	-
Ruoli dei partecipanti	variabili
Movimento	+/-
Manipolazione oggetti	+
Multiattività	+
Quantità	
Eventi	5
Durata	6 h 05'



Preparazione di cibo	
Parametri	
Numero di partecipanti	4
Istituzionalità	-
Ruoli dei partecipanti	variabili
Movimento	+
Manipolazione oggetti	+
Multiattività	++
Quantità	
Eventi	2
Durata	1 h 43'

- 1 SA \*io questo non ho ancora capito dove va Δbuttato. (0.6) Δ  
\*gz carton-->  
Δturns it-----Δ
- 2 SA Δperché qua c'è scritto tetrapack\*Δ  
-->\*  
Δpoints the bottom and shows it---Δ  
fig #fig.1
- 3 SA \*ma il tetrapack va buttato nel sacchetto normale  
\*gz at ME-->
- 4 perché non c'è un::\*  
-->\*
- 5 ME anche perché \$dentro: c'è\$ [l'alluminio.]  
\$points carton\$
- 6 CR [ma c'è un] disegno del +sacchetto. Δ=guarda:. Δ  
+points the carton-->  
Δturns it-Δ
- 7 \* (1.45)
- 8 CR Δgira. Δ+  
-->+turn gest-->  
Δturns the cartonΔ
- 9 (0.38)%  
me %gz carton-->

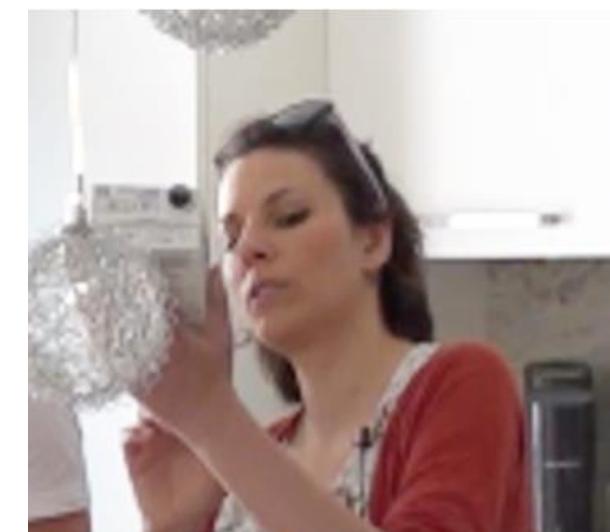


fig. 1: SA

- 10 SA      no Δ[c'è scritto] tetrapack.Δ  
               Δturns the carton-----Δ
- 11 CR      [gira.] +  
               [turn it]  
               --> +
- 12           (0.23)
- 13 CR      \*+sacchetto.+  
               +points the image on the side of  
               sa      -->\*gz pointed side-->>  
               fig      #fig.2
- 14 ME      sacchetto normale.  
               \$si ma perché dentro% c'è\$ l'alu  
               -->%  
               \$hand twd the carton-----\$
- 15 /  
               quindi non puoi neanche metterlo nella carta.



fig 2.: CR, SA, ME



## Tutoring in architettura

Parametri

Numero di partecipanti

2-3

Istituzionalità

+

Ruoli dei partecipanti

asimmetrici

Movimento

+/-

Manipolazione oggetti

+

Multiattività

+

Quantità

Eventi

4

Durata

4 h 41'



<b>Lezioni di teatro, musica, restauro</b>	
<b>Parametri</b>	
<b>Numero di partecipanti</b>	<b>6-11</b>
<b>Istituzionalità</b>	<b>+</b>
<b>Ruoli dei partecipanti</b>	<b>asimmetrici</b>
<b>Movimento</b>	<b>+</b>
<b>Manipolazione oggetti</b>	<b>+</b>
<b>Multiattività</b>	<b>++</b>
<b>Quantità</b>	
<b>Eventi</b>	<b>3</b>
<b>Durata</b>	<b>4 h 51'</b>



<b>Lezioni di cultura generale / glottodidattica</b>	
<b>Parametri</b>	
<b>Numero di partecipanti</b>	8-19
<b>Istituzionalità</b>	+
<b>Ruoli dei partecipanti</b>	asimmetrici
<b>Movimento</b>	+/-
<b>Manipolazione oggetti</b>	+/-
<b>Multiattività</b>	-
<b>Quantità</b>	
<b>Eventi</b>	5
<b>Durata</b>	6 h 05'

- 1 Giuseppe hm; questo non c'è più; (1.24) †no (.) e non  
2 potresti usare-QUESTo SPAzio inTERno,  
3 (2.75)
- 4 Clara eh <> avevo> fatto la prova con tutte  
5 le scale che vanno: su li-  
6 Giuseppe ma non funziona,  
7 Clara no; (0.64) cioè il professore <> aveva>  
8 Giuseppe [eh; (. )]  
9 Clara [detto che non:] (-) <> dim> non era convincente;>  
10 (4.38)





Interviste	
Parametri	
Numero di partecipanti	2
Istituzionalità	+
Ruoli dei partecipanti	asimmetrici
Movimento	-
Manipolazione oggetti	-
Multiattività	-
Quantità	
Eventi	6
Durata	3 h 42'

RICERCATORE1

e: però l~=la tua tesi era più di di:  
(0.94)

Gabriella

sì [la:]

RICERCATORE1

[sullo] studio del dialetto ho sentito

Gabriella

sì=sì=sì=sì=sì

RICERCATORE1

[(hai) fatto intervi]ste:;

Gabriella

[sì era:] (0.13) sì;



# 4. Parlanti del TIGR: chi sono?

## Campagna di reclutamento fine 2020 – inizio 2022

- Mailing list interne all'università
- Social media
- Contatti personali
- Presa di contatto diretta con istituzioni

[https://usi.qualtrics.com/jfe/form/SV\\_1B4aGWceuGAC5x4](https://usi.qualtrics.com/jfe/form/SV_1B4aGWceuGAC5x4)

◆ Pinned Tweet



Infinita @ItalInfin · 10s

Parlanti cercansi! Per documentare l'italiano quotidiano parlato nella #Svizzeraitaliana raccogliamo il #CorpusTIGR, che comprende conversazioni videoregistrate 🎥, trascritte abcd e anonimizzate ?. Interessata/o? Per più info e per partecipare all'iniziativa: [bit.ly/CorpusTIGR](http://bit.ly/CorpusTIGR)



The categorization  
of common  
sources in  
face-to-face  
interaction.  
A study based  
on the TIGR corpus  
of spoken Italian

### Chi siamo

*Infinita* è un progetto di ricerca in linguistica italiana condotto presso l'Università della Svizzera italiana e finanziato dal Fondo Nazionale Svizzero (sussidio n. 192771).

### Il corpus TIGR

Nel quadro di Infinita raccogliamo il corpus TIGR, una risorsa innovativa e preziosa per lo studio dell'italiano parlato in Ticino e nel Grigioni. Il corpus comprenderà conversazioni in diverse situazioni di interazione spontanea. Ogni conversazione verrà videoregistrata con dispositivi facilmente trasportabili (2 videocamere, microfoni senza fili), trascritta e anonimizzata.

Per fare questo, **abbiamo bisogno di te!** Ti spiegheremo tutto in questo breve modulo.



## Dichiarazioni di consenso informato: esigenze della ricerca, esigenze legali e questioni etiche

- Modello USI e contatti con Commissione etica e Servizio legale
- Riferimenti legislativi e linee guida
- Prevedere fin dall'inizio la pubblicazione dei dati (video) richiede una riflessione sui contenuti delle dichiarazioni:

Come gestire la **de-identificazione** dei parlanti? come descrivere in modo informativo ma non limitante il futuro **riuso dei dati**? Quali impegni possiamo prendere come ricercatori?

- Legge federale sulla protezione dei dati (LDP) del 25 settembre 2020 (Stato 1 settembre 2023). <https://www.fedlex.admin.ch/eli/cc/2022/491/it>
- Kruegel, S. (2019). The informed consent as legal and ethical basis of research data production. FORS Guide No. 05, Version 1.0. Lausanne: Swiss Centre of Expertise in the Social Sciences FORS. [doi:10.24449/FG-2019-00005](https://doi.org/10.24449/FG-2019-00005)
- Profazi, N., Miecznikowski, J. (2023). Social interaction is among people. Legal, technical, and ethical explorations about personal information and its removal in talk-in-interaction as data. <https://www.chord-talk-in-interaction.usi.ch/news/feeds/36387>

## Dichiarazioni di consenso informato: esigenze della ricerca, esigenze legali e questioni etiche

### A. Documento informativo

1. Introduzione
2. Descrizione progetto di ricerca per pubblico non specialista
3. Descrizione dell'evento e modalità di video-registrazione
4. **Confidenzialità e protezione dei dati**
5. **Conservazione e utilizzo dei dati**
6. Diritti dei partecipanti allo studio
7. Contatti

### B. Consenso informato

Scelta tra diversi misure di de-identificazione

- **Default** > pseudonimi nella trascrizione, silenzi nella traccia audio, misure di de-identificazione nei metadati
- **Misure supplementari** > alterazione della voce nella traccia audio (6 richieste), applicazione di filtri nella traccia video (15 richieste)

## Questionario sociolinguistico

ID partecipante (a cura dei ricercatori) \_\_\_\_\_

7. Professione \_\_\_\_\_

1. Età

\_\_\_\_\_

8. Istruzione

2. Sesso

M

F

Scuole elementari

Licenza di scuola media

Formazione professionale

Diploma di scuola media superiore

Laurea triennale

Laurea magistrale

Dottorato

Altro \_\_\_\_\_

3. Comune dell'istruzione primaria

\_\_\_\_\_

4. Comune di residenza

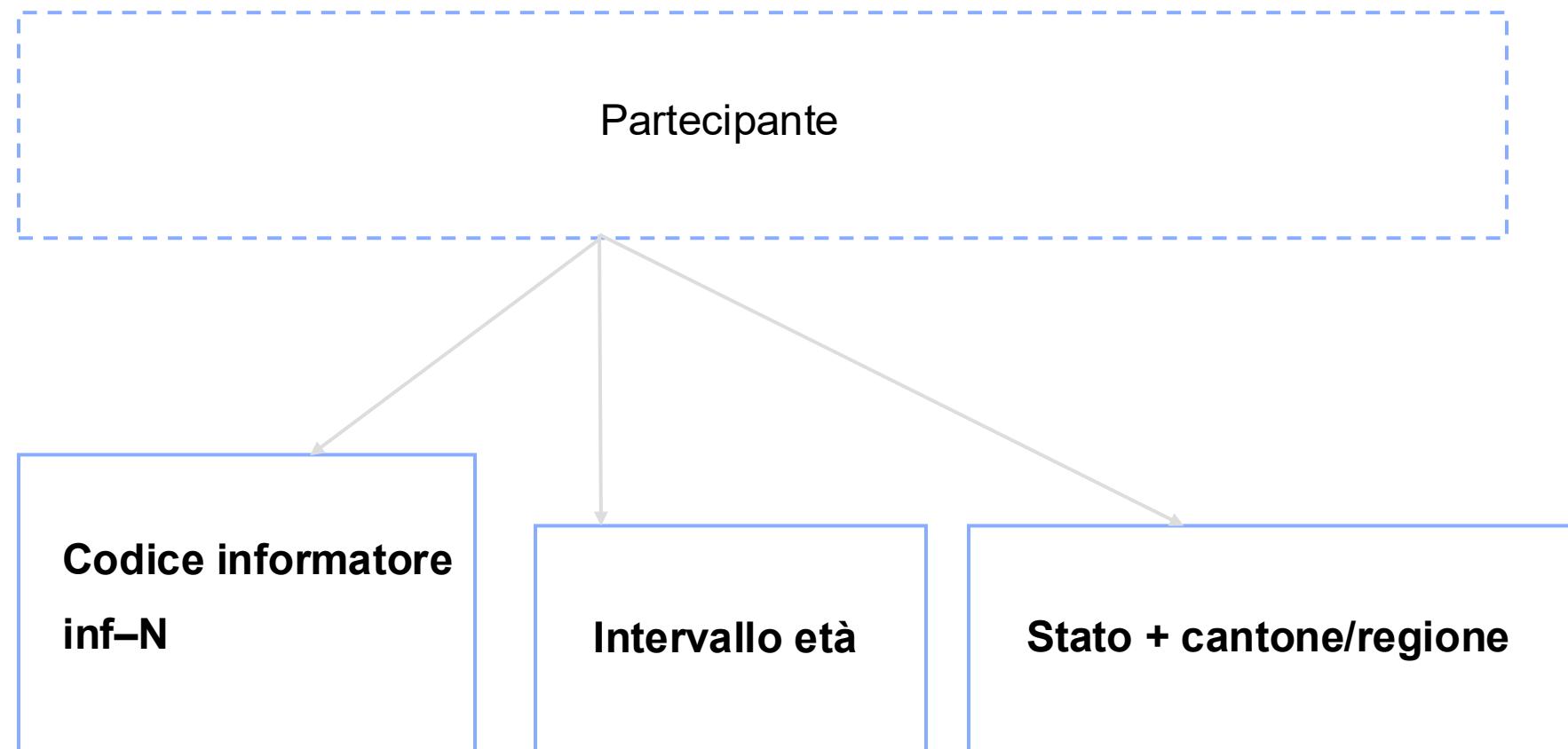
\_\_\_\_\_

5. Comune di lavoro o studio in Svizzera

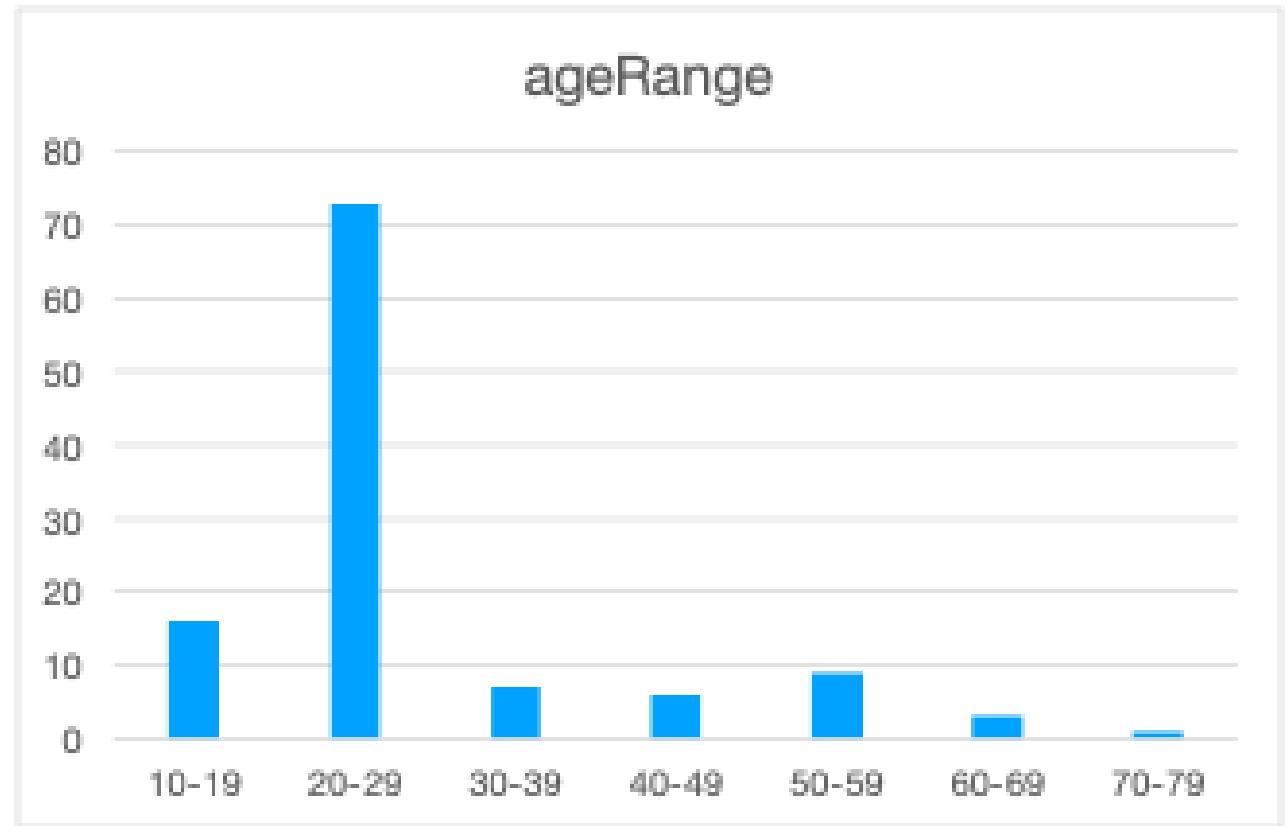
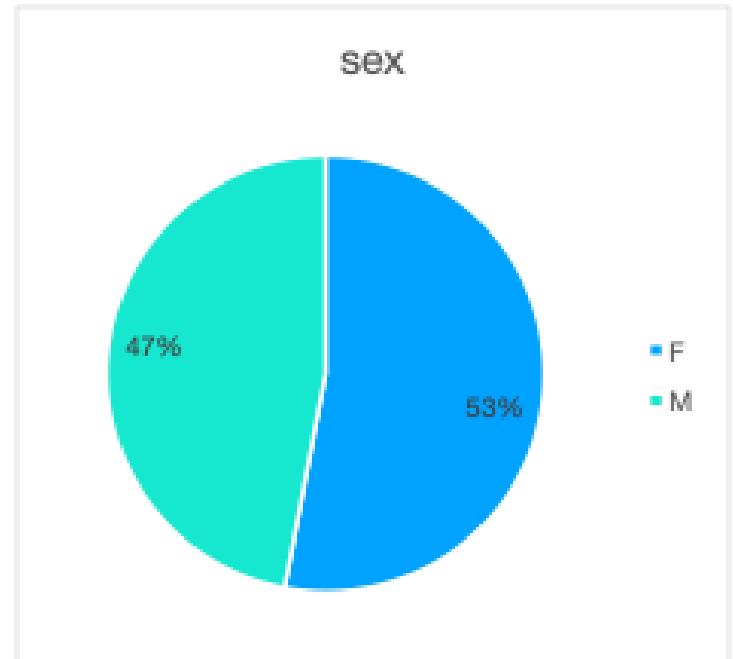
\_\_\_\_\_

6. Lingue conosciute

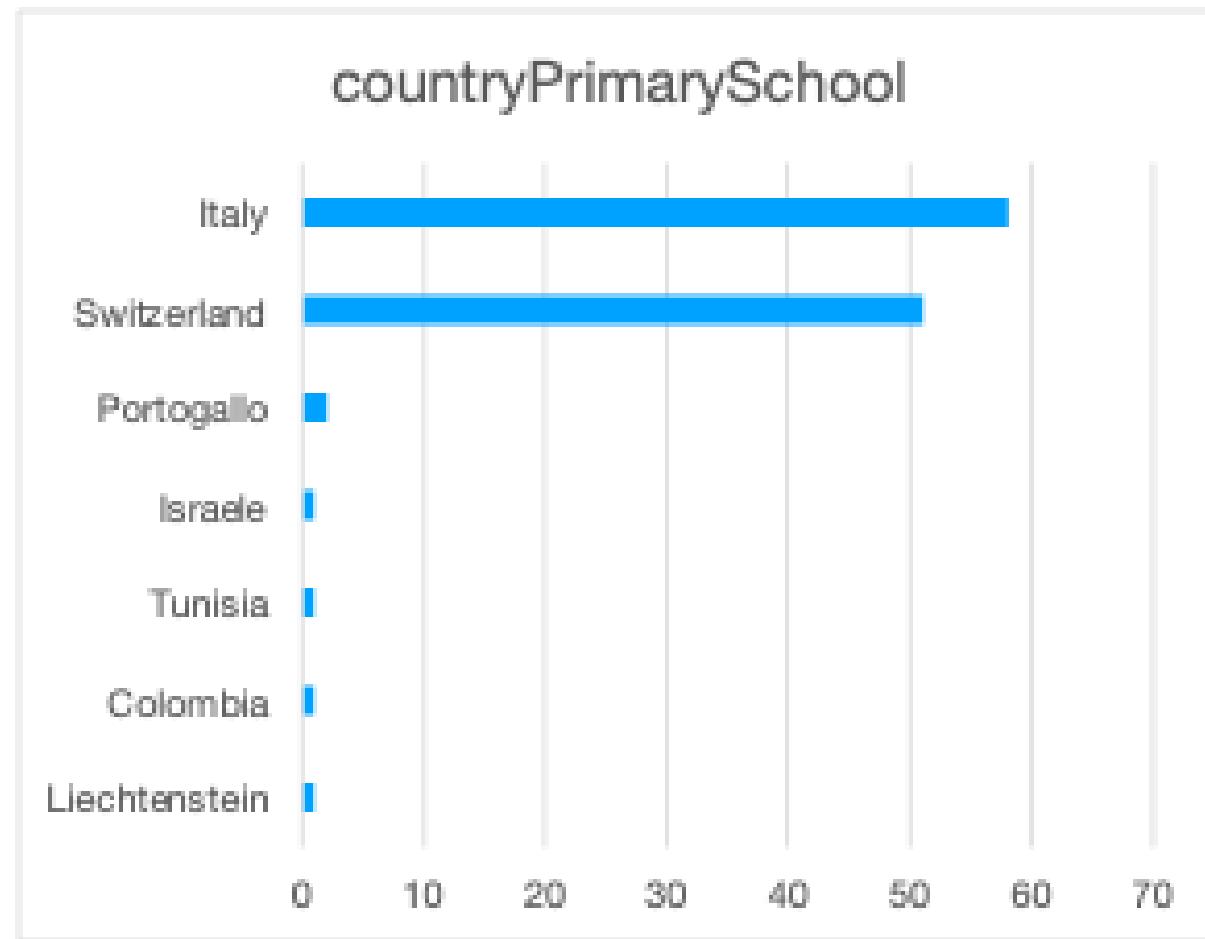
## Misure di de-identificazione



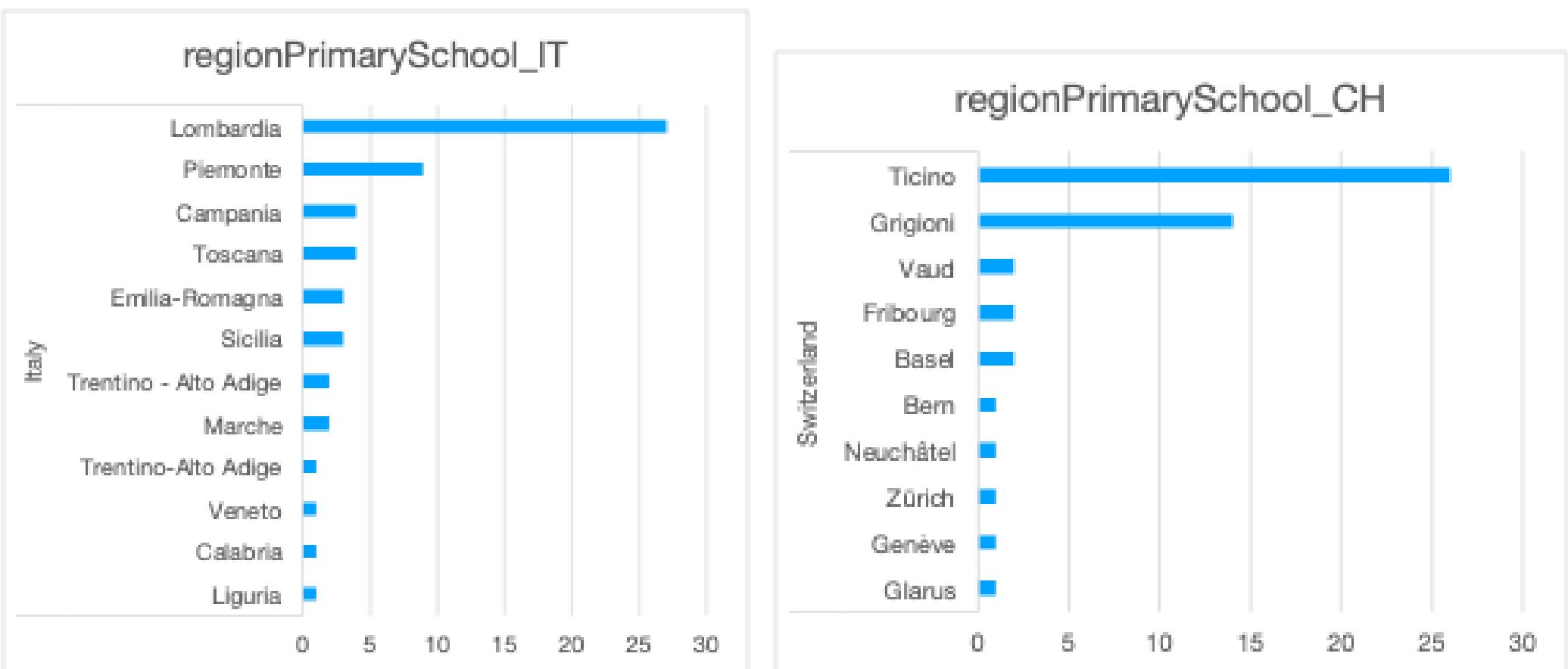
## Dati sociolinguistici: genere e età



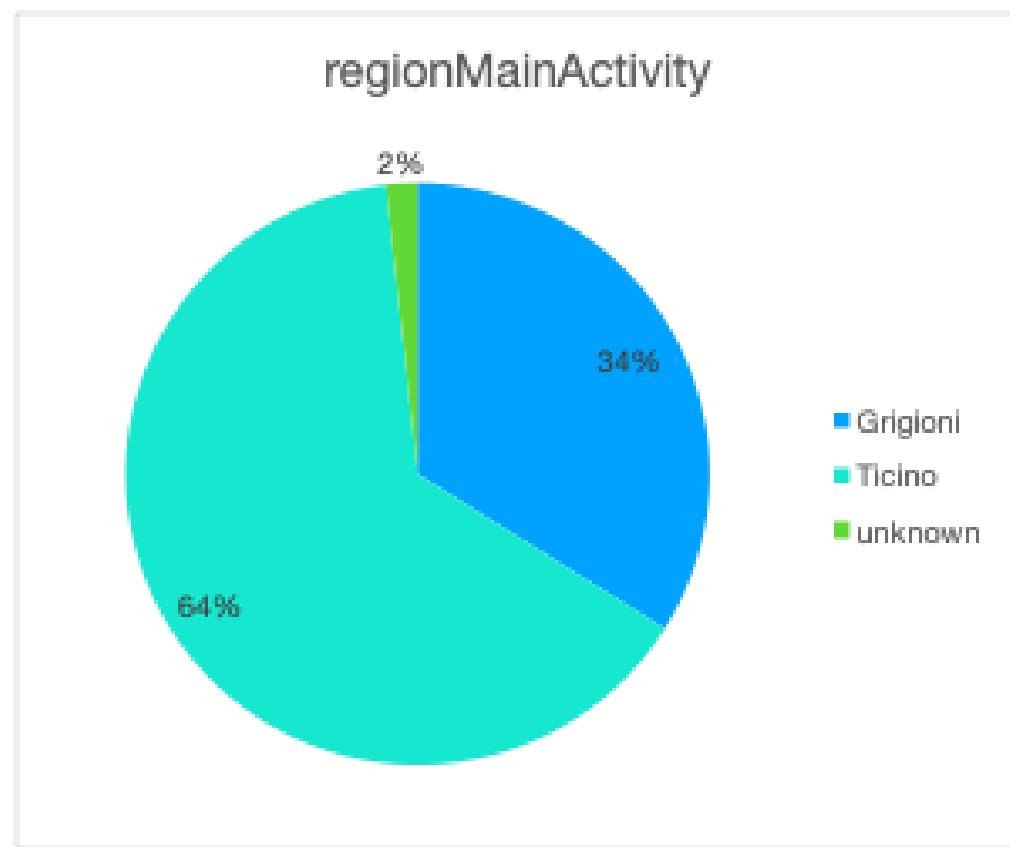
## Dati sociolinguistici: origine dei parlanti



## Dati sociolinguistici: origine dei parlanti

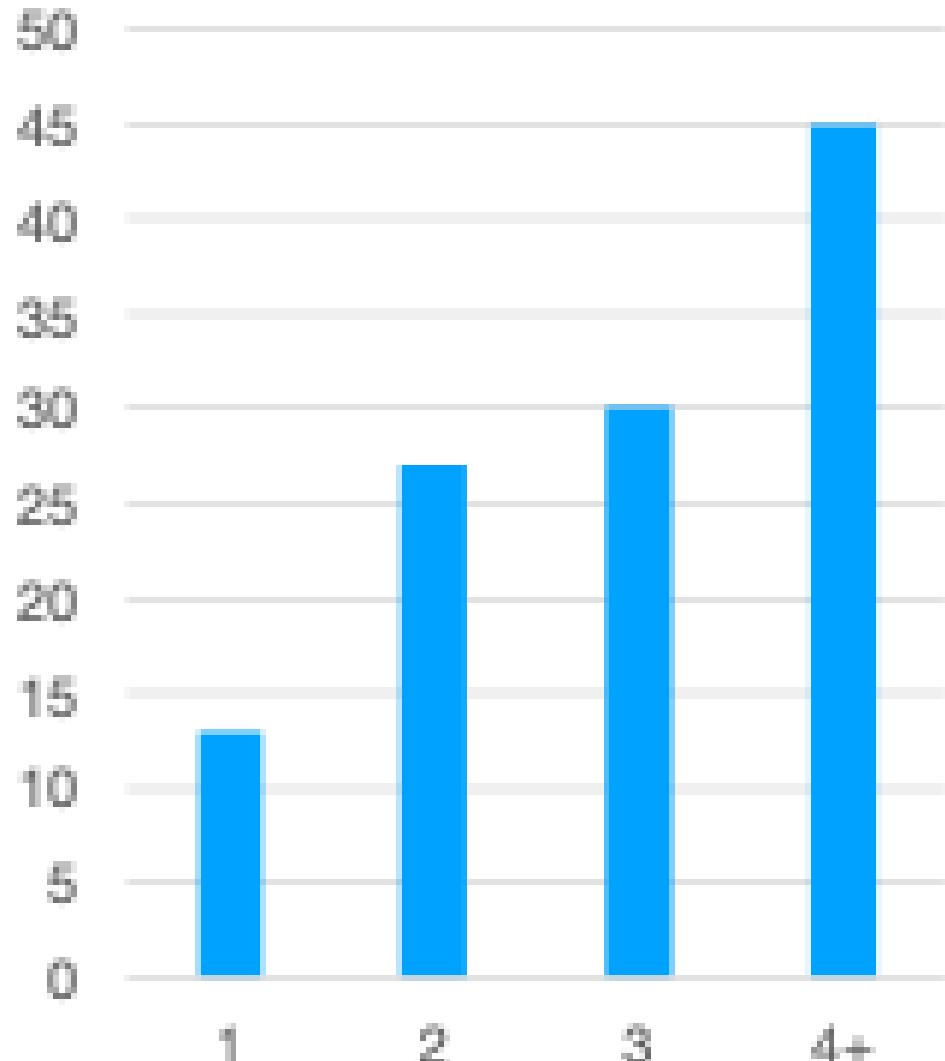


## Dati sociolinguistici: luogo di lavoro o studio in Svizzera

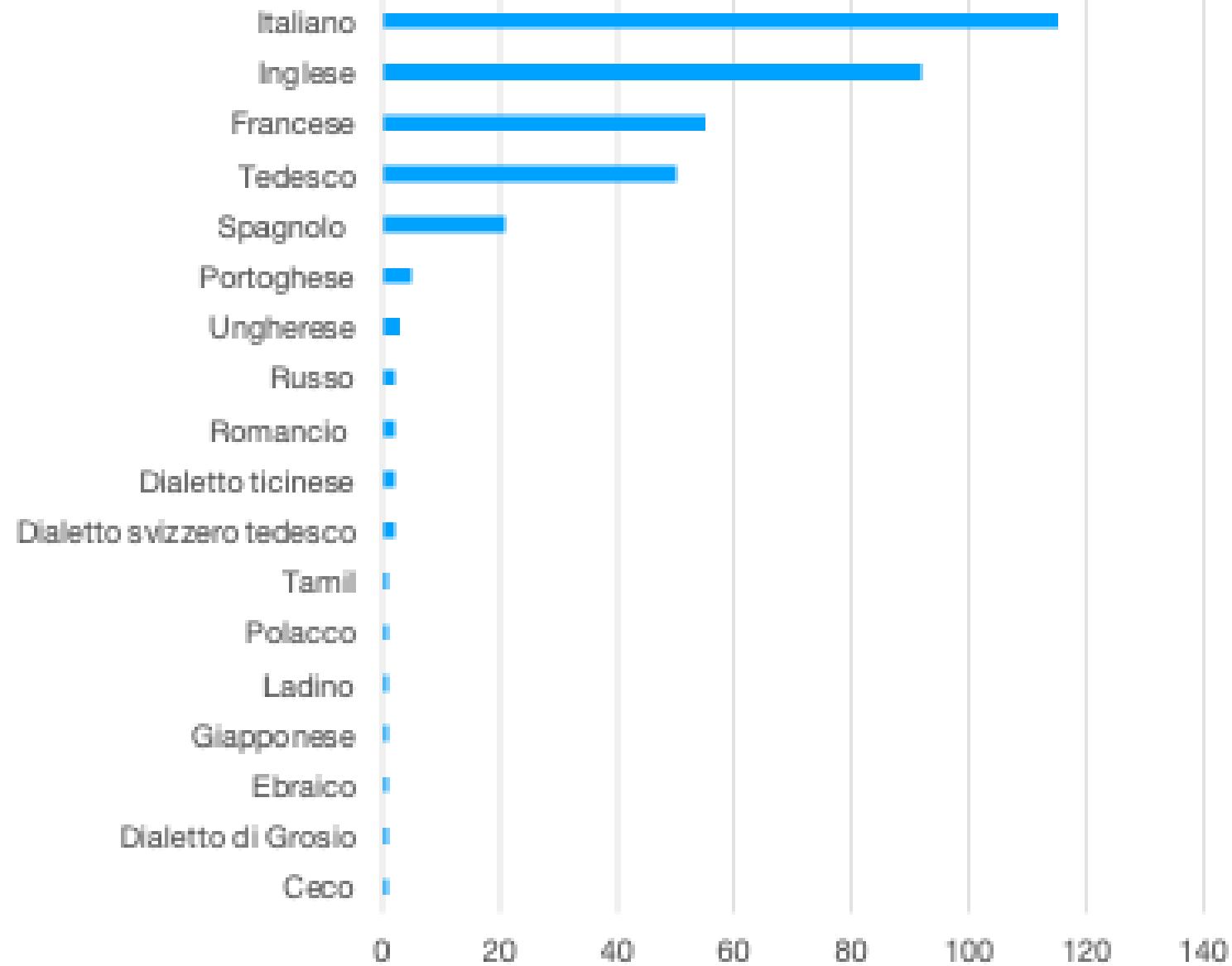


## Dati sociolinguistici: repertorio linguistico

### languageRepertoire



### languages



## Dati sociolinguistici: istruzione e professione

### education

Secondary 2 general education



Professioni prevalenti: studenti universitari; apprendisti in formazione professionale; insegnanti

Secondary 2 vocational  
education and training

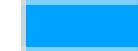


B.A.



Altre professioni: impiegato/a; consulente bancario, pedagogico, informatico; ingegnere; architetto/a; muratore; cuoco/a; infermiere/;, pensionato/a

M.A.



PhD



Primary school



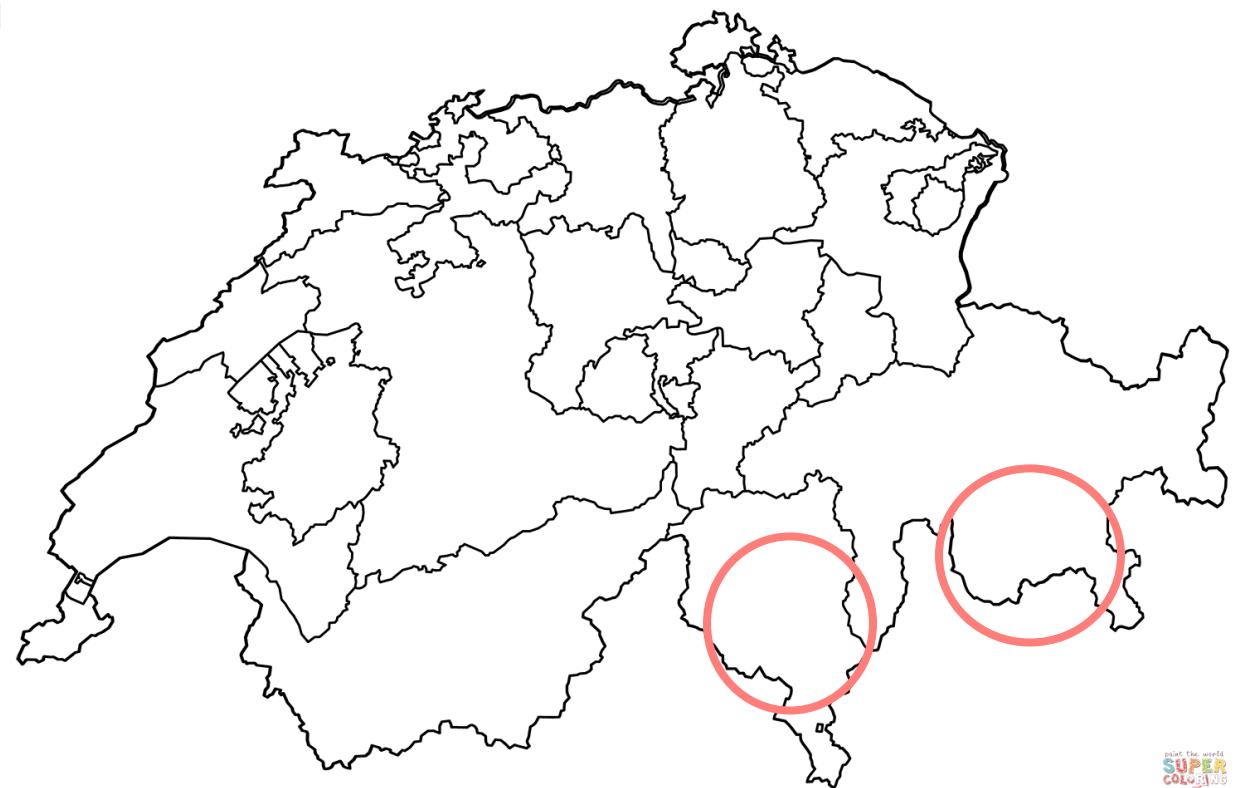
unknown



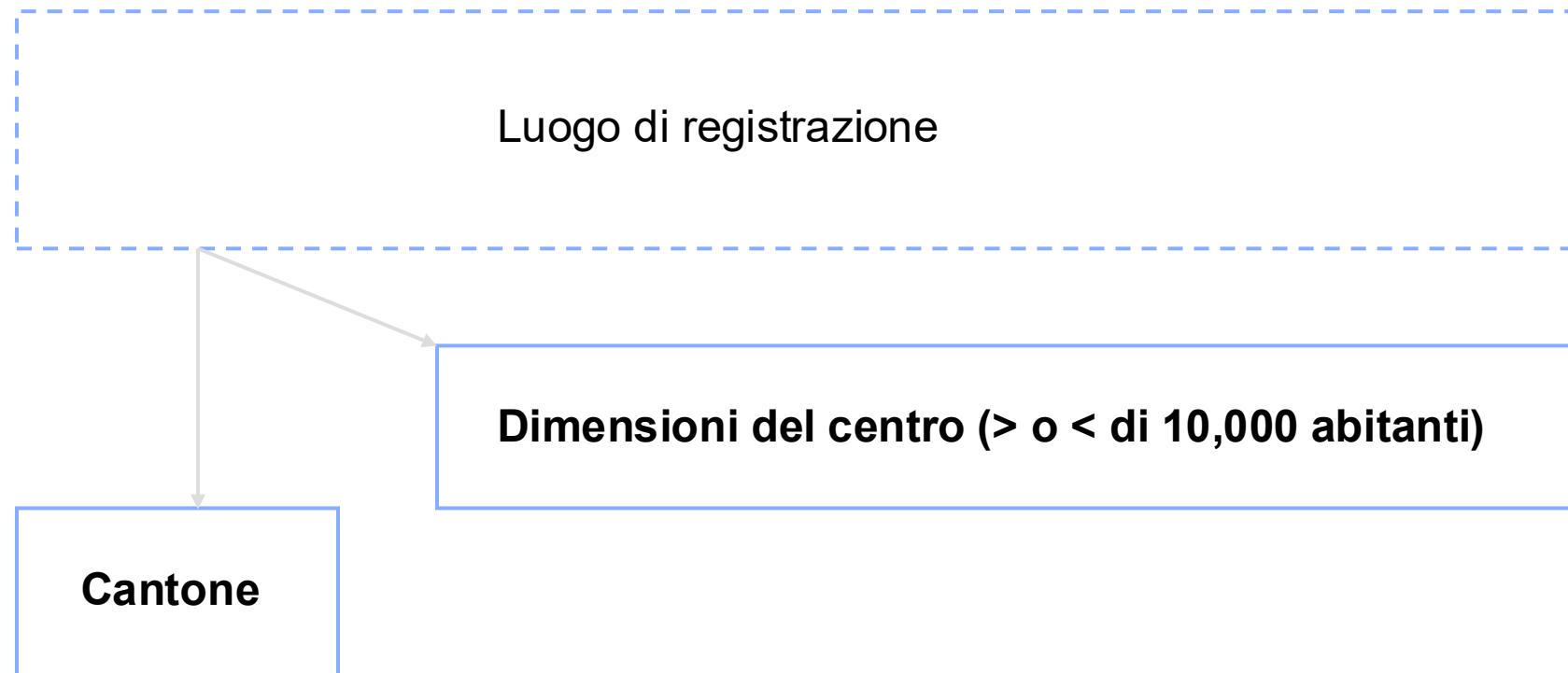
# 5. I luoghi del TIGR

## Criteri di scelta dei luoghi

- **Ticino e Grigioni**
- Le località sono determinate in modo opportunistico in base alla disponibili dei parlanti, no restrizioni imposte. Il corpus non è disegnato per documentare la variazione diatopica dell'Italiano parlato in Svizzera
- La presa di contatto ad hoc con le istituzioni di formazione ha tuttavia permesso di ampliare la copertura geografica del corpus.



## Misure di de-identificazione



Ticino

9 eventi in 7 piccoli centri

9 eventi in 2 grandi centri

---

**18 eventi in 9 località**



Grigioni

4 eventi in 1 piccolo centro

1 evento in 1 grande centro

---

**5 eventi in 2 località**



# 6. Impostazione tecnica delle registrazioni



- 2 Sony HXR-NX80//C camcorders
- 2-4 Tentacle Track E pocket audio recorders with clip-on microphones
- 1-2 additional microphones for larger groups
- Synchronisation of devices before recording through Tentacle timecode generators (digital TC in pocket recorders, acoustic TC in timecode generators for camcorders)

# 7. Post-produzione dei dati audio e video

## Tentacle Timecode Tool

- Conversione del timecode acustico in timecode digitale

## Adobe Premiere

- Sincronizzazione di tutte le tracce in base al timecode digitale
- Taglio per ottenere tracce di uguale lunghezza
- De-identificazione dei volti (ca. 10% dei partecipanti): effetto video Gaussian blur
- Masterizzazione delle tracce audio e creazione di un video integrato a schermo diviso

## ELAN/Praat

- Per ogni trascrizione, esportazione da ELAN di una lista di intervalli contenenti nomi, marcatori su un tier apposito
- Sostituzione di questi intervalli con silenzi in Praat grazie a uno script (© Francesco Cangemi).

De-identificazione delle voci?? (pochi casi)

# 8. Trascrizione

# Software ELAN

ELAN 6.6 - ev-2-REV-EB-09.08.2023.eaf

File Edit Annotation Tier Type Search View Options Window Help

Selection: 00:00:00.000 - 00:00:00.000 0

Volume:

- audio camera mic ev2.wav (Mute Solo)
- audio R791 ev2.wav (Mute Solo)
- audio R793 ev2.wav (Mute Solo)
- audio R796 ev2.wav (Mute Solo)
- camera mic ev2 no audio ridotto.mp4 (Mute Solo)
- camera no mic ev2 no audio ridotto.mp4 (Mute Solo)

Selection Mode Loop Mode

audio camera ... 00:00:00.000 - 00:00:17.000

AMBENT\_NOISES 00:00:00.000 - 00:00:17.000

CAROLA 796 00:00:00.000 - 00:00:17.000

CARLA 793 00:00:00.000 - 00:00:17.000

ALESSIO 791 00:00:00.000 - 00:00:17.000

ANONYMIZATION 00:00:00.000 - 00:00:17.000

vado a vedere a che punto è la super gentile. ((laughs)) [si. h] [graz]

dai sono stato genTII cè <<len> non avevo:: detto nie ecco. (-) brava carla che secondo me <<all>è l'unica che si è ricordata di chiedergli se voleva da bere. > ((laughs)) \*h [d~] dopo venti minuti che [era qua a sistemare]

NA

## Un lavoro di squadra

- La trascrizione è stata finora svolta da tre studentesse assistenti impiegate in periodi diversi
- Attualmente collaboriamo con una studentessa e uno studente assistenti per completare la trascrizione
- Il team del progetto lavora sulla revisione delle trascrizioni, sull'esportazione e la formattazione

## Convenzioni di trascrizione

- **Sistema GAT2 (Gesprächsanalytische Transkriptionssystem)**
- Sviluppato più tardi rispetto a Jefferson, già funzionale a una trascrizione tramite software e machine-readable
- Integrato in editor digitali (per esempio EXMARaLDA)
- Livelli di granularità: minimal transcript, basic transcript, **fine transcript**
  - + Lista di lemmi per uniformare la grafia delle interiezioni
  - + Trascrizione del raddoppiamento fonosintattico (ex., *va bbene, e nniente*)
  - + Uso del simbolo ~ per interruzioni udibili di una parola (in particolare, glottal stop), per esempio in esitazioni e ripartenze

Selting, M., Auer, P., Barth-Weingarten, D., Bergmann, J. ö, Bergmann, P., Birkner, K., Couper-Kuhlen, E., Deppermann, A., Gilles, P., Günthner, S., Hartung, M., Kern, F., Mertzlufft, C., Meyer, C., Morek, M., Oberzaucher, F., Peters, J. ö, Quasthoff, U., Schütte, W., & Uhmann, S. (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.

### Sequential structure

[ ]	overlap and simultaneous talk
[ ]	
=	fast, immediate continuation with a new turn or segment (latching)

### In- and outbreaks

°h / h°	in- / outbreaks of appr. 0.2-0.5 sec. duration
°hh / hh°	in- / outbreaks of appr. 0.5-0.8 sec. duration
°hhh / hhh°	in- / outbreaks of appr. 0.8-1.0 sec. duration

### Pauses

(.)	micro pause, estimated, up to 0.2 sec. duration appr.
(-)	short estimated pause of appr. 0.2-0.5 sec. duration
(--)	intermediary estimated pause of appr. 0.5-0.8 sec. duration
(---)	longer estimated pause of appr. 0.8-1.0 sec. duration
(0.5) / (2.0)	measured pause of appr. 0.5 / 2.0 sec. duration (to tenth of a second)

Other segmental conventions

:	lengthening, by about 0.2-0.5 sec.
::	lengthening, by about 0.5-0.8 sec.
:::	lengthening, by about 0.8-1.0 sec.
?	cut-off by glottal closure
and_uh	cliticizations within units
uh, uhm, etc.	hesitation markers, so-called "filled pauses"

Laughter and crying

haha, hehe, hihi	syllabic laughter
((laughs)), ((cries))	description of laughter and crying
<<laughing> >	laughter particles accompanying speech with indication of scope
<<:-> so>	smile voice

Continuers

hm, yes, no, yeah	monosyllabic tokens
hm_hm, ye_es, no_o	bi-syllabic tokens
?hm?hm	with glottal closure, often negating

Accentuation

SYLlable	focus accent
sYllable	secondary accent
!SYL!lable	extra strong accent

Final pitch movements of intonation phrases

?	rising to high
,	rising to mid
-	level
;	falling to mid
.	falling to low

Pitch jumps

↑	smaller pitch upstep
↓	smaller pitch downstep
↑↑	larger pitch upstep
↓↓	larger pitch downstep

Changes in pitch register

<<1>>	lower pitch register
<<h>>	higher pitch register

Loudness and tempo changes, with scope

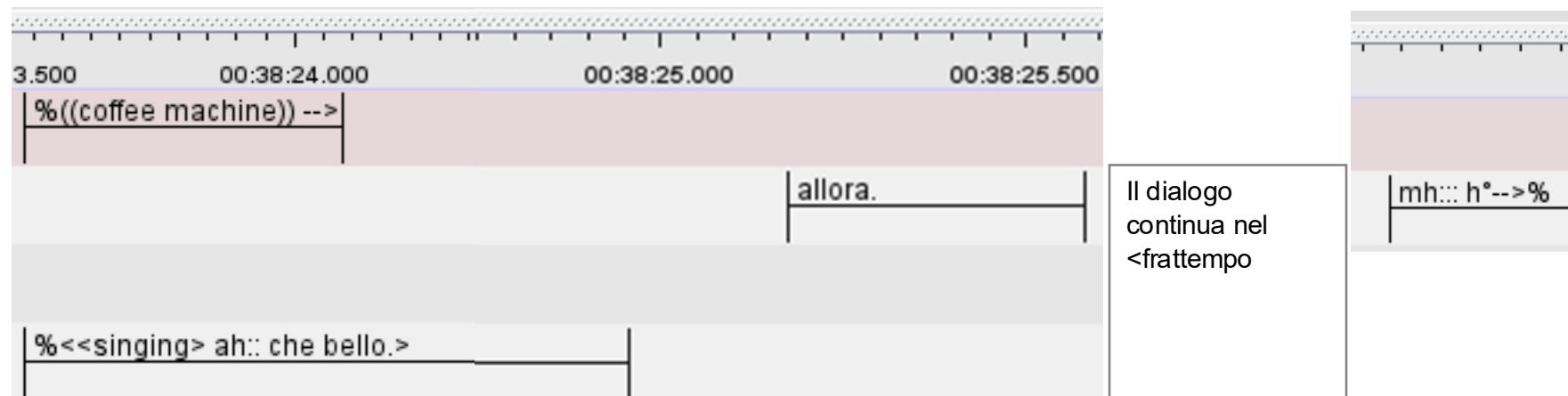
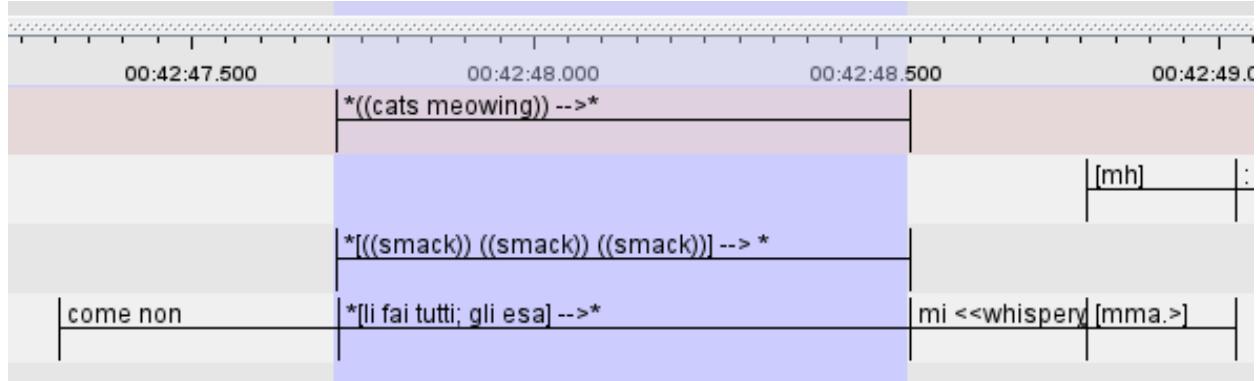
<<f>	>	forte, loud
<<ff>	>	fortissimo, very loud
<<p>	>	piano, soft
<<pp>	>	pianissimo, very soft
<<all>	>	allegro, fast
<<len>	>	lento, slow
<<cresc>	>	crescendo, increasingly louder
<<dim>	>	diminuendo, increasingly softer
<<acc>	>	accelerando, increasingly faster
<<rall>	>	rallentando, increasingly slower

Changes in voice quality and articulation, with scope

<<creaky>	>	glottalized
<<whispery>	>	change in voice quality as stated
<u>Other conventions</u>		
<<surprised>	>	interpretive comment with indication of scope
((coughs))		non-verbal vocal actions and events
<<coughing>	>	...with indication of scope
( )		unintelligible passage
(xxx), (xxx xxx)		one or two unintelligible syllables
(may i)		assumed wording
(may i say/let us say)		possible alternatives
((unintelligible, appr. 3 sec))		unintelligible passage with indication of duration
((...))		omission in transcript
-->		refers to a line of transcript relevant in the argument

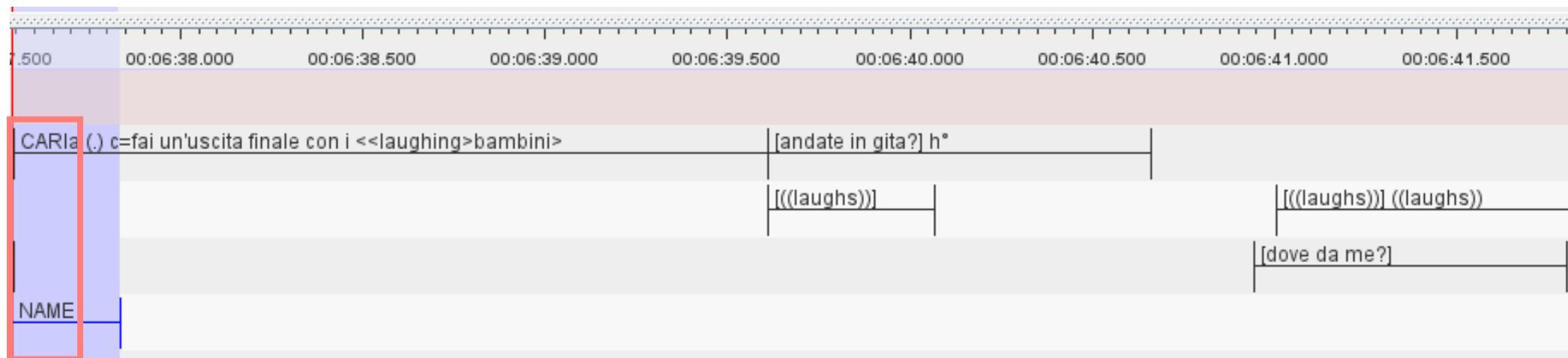
Per il tier <ambient noises>, un sistema di ancoraggio al testo ispirato a quello proposto da Mondada per la trascrizione multimodale.

Mondada, L. (2018). Multiple Temporalities of Language and Body in Interaction: Challenges for Transcribing Multimodality, *Research on Language and Social Interaction*, 51:1, 85-106.



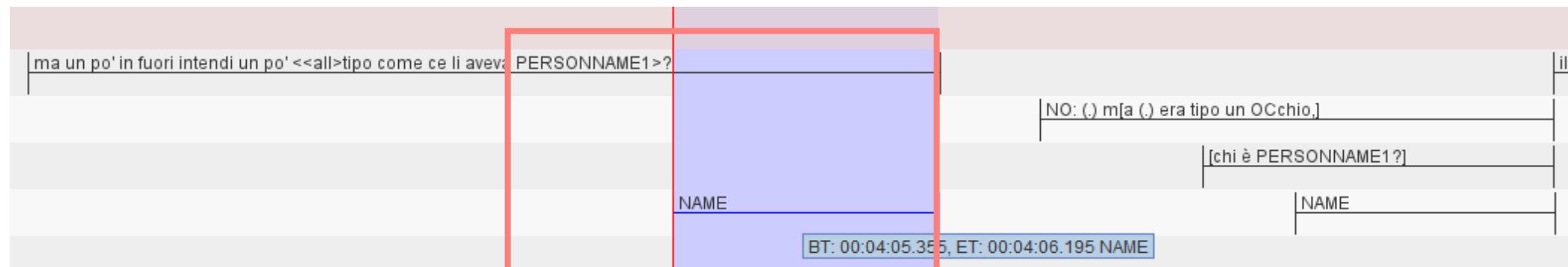
## Misure di de-identificazione

- Attribuzione di **pseudonimi** solo ai partecipanti all'interazione in corso



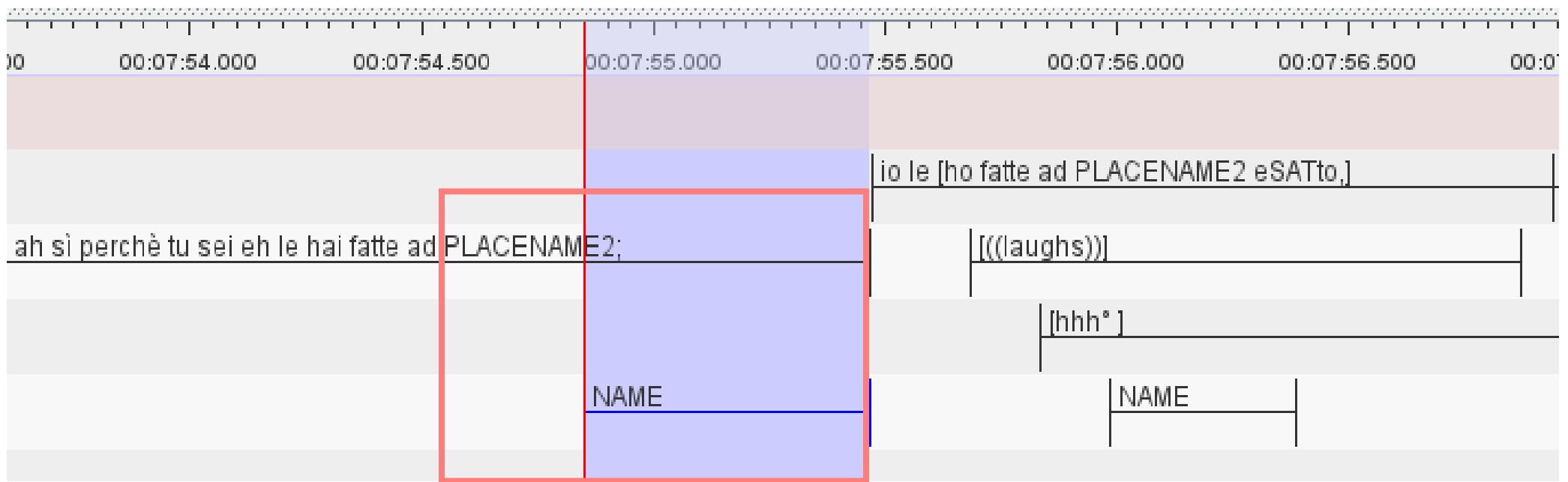
## Misure di de-identificazione

- Nomi di altre persone > formato PERSONNAME



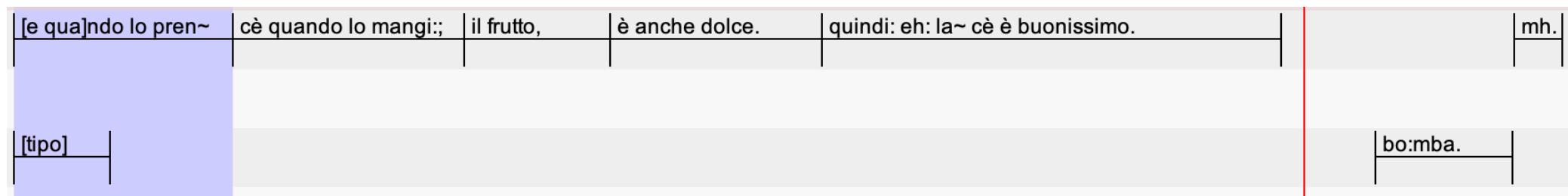
## Misure di de-identificazione

- Nomi di luoghi, istituzioni...> formato PLACENAME, INSTITUTIONNAME

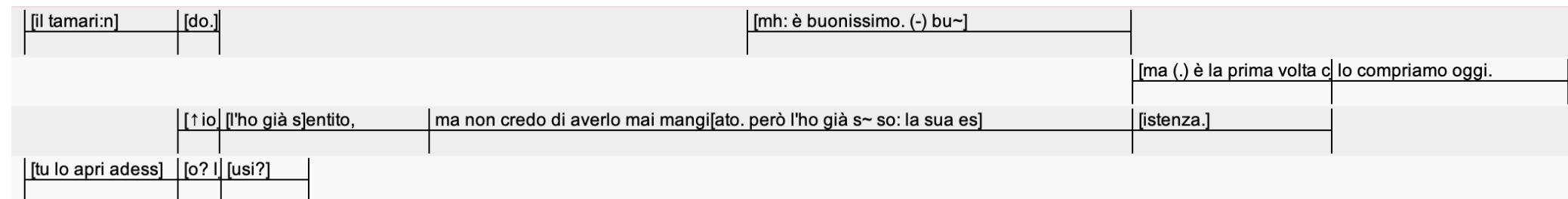


## Segmentazione in ELAN

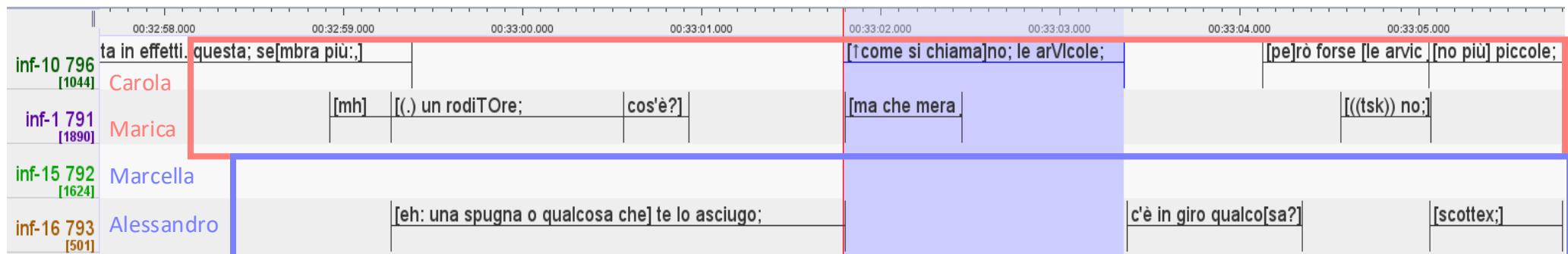
- Il segmento non rappresenta necessariamente un'unità significativa sul piano linguistico
- Segmentazione funzionale alla trascrizione accurata, per esempio nel caso di sovrapposizioni multiple



VS.



## Il problema degli schismi



# 9. La derivazione di trascrizioni in formato testo in stile ‘dialogo teatrale’

ev-4\_SK.txt - Notepad  
File Edit Format View Help

Marcella <>laughing> hm (-) hm hm [hm>)])] [neanche] [qui?]  
Carola [vengo] [con te;]  
(0.35)  
Marica vieni [con me?]  
Marcella [io me]tto; (.) QUE[Sto] ((TC)) in centro,  
--TIMECODE-- 00:01:00.177  
Carola [hm\_hm]  
Marcella (-) togliamo la  
Marica gna[m;]  
Marcella [can]delina, che (.) tan\*to (-) fa\* troppo caldo  
per accenderla.  
AMBIENT\_NOISES \*((rumore di piatti))\*(3.79)  
Marica <>whispery> ecco.>  
(1.03)  
AMBIENT\_NOISES ((TC)) §((rumore di acqua di rubinetto))-->  
--TIMECODE-- 00:01:11.040  
Marica §oh scusa[mi]  
Marcella [come sei ales][sandro?]  
Carola [figura][ti.]  
Marica [ti son pas][sata avan]ti,

- Trascrizioni in stile ‘dialogo teatrale’ e formato TXT:
  - approssimazione grafica dell’organizzazione della conversazione in turni e buona base per un’analisi sequenziale
  - input per l’annotazione delle fonti di informazione all’interno del progetto InfinIta, eseguita nel programma Inception
  - text/plain: un formato semplice e interoperabile
- Esportazione da ELAN: in formato testo “tradizionale”, con o senza timecode (tempi iniziali e finali dei segmenti), formattazione automatica dei turni (diverse opzioni)
- Problemi: quantità di timecode, ordine lineare non sempre intuitivo
- Soluzione: una procedura semi-automatica (script Python) per creare trascrizioni TXT
  - esportazione da ELAN come trascrizione in formato “tradizionale” e con allineamento verticale delle parentesi quadre
  - eliminazione dei timecode tranne tempi iniziali all’incirca ogni 10 secondi, collocazione di ancora “((TC))” nel testo
  - eliminazione dei nomi di parlanti superflui, concatenazione dei segmenti
  - formattazione manuale dei passi con sovrapposizioni di più di due parlanti.

ev-4\_SK.txt - Notepad  
File Edit Format View Help

Marcella <<laughing> hm (-) hm hm [hm>)])] [neanche] [qui?]  
Carola [vengo] [con te;]  
(0.35)  
Marica vieni [con me?]  
Marcella [io me]tto; (.) QUE[Sto] ((TC)) in centro,  
--TIMECODE-- 00:01:00.177  
Carola [hm\_hm]  
Marcella (-) togliamo la  
Marica gna[m;]  
Marcella [can]delina, che (.) tan\*to (-) fa\* troppo caldo  
per accenderla.  
AMBIENT\_NOISES \*((rumore di piatti))  
(3.79)  
Marica <<whispery> ecco.>  
(1.03)  
AMBIENT\_NOISES ((TC)) §((rumore di acqua di rubinetto))-->  
--TIMECODE-- 00:01.11.040  
Marica §oh scusa[mi]  
Marcella [come sei ales][sandro?]  
Carola [figura][ti.]  
Marica [ti son pas][sata avan]ti,

# 10. Verso una versione tokenizzata delle trascrizioni

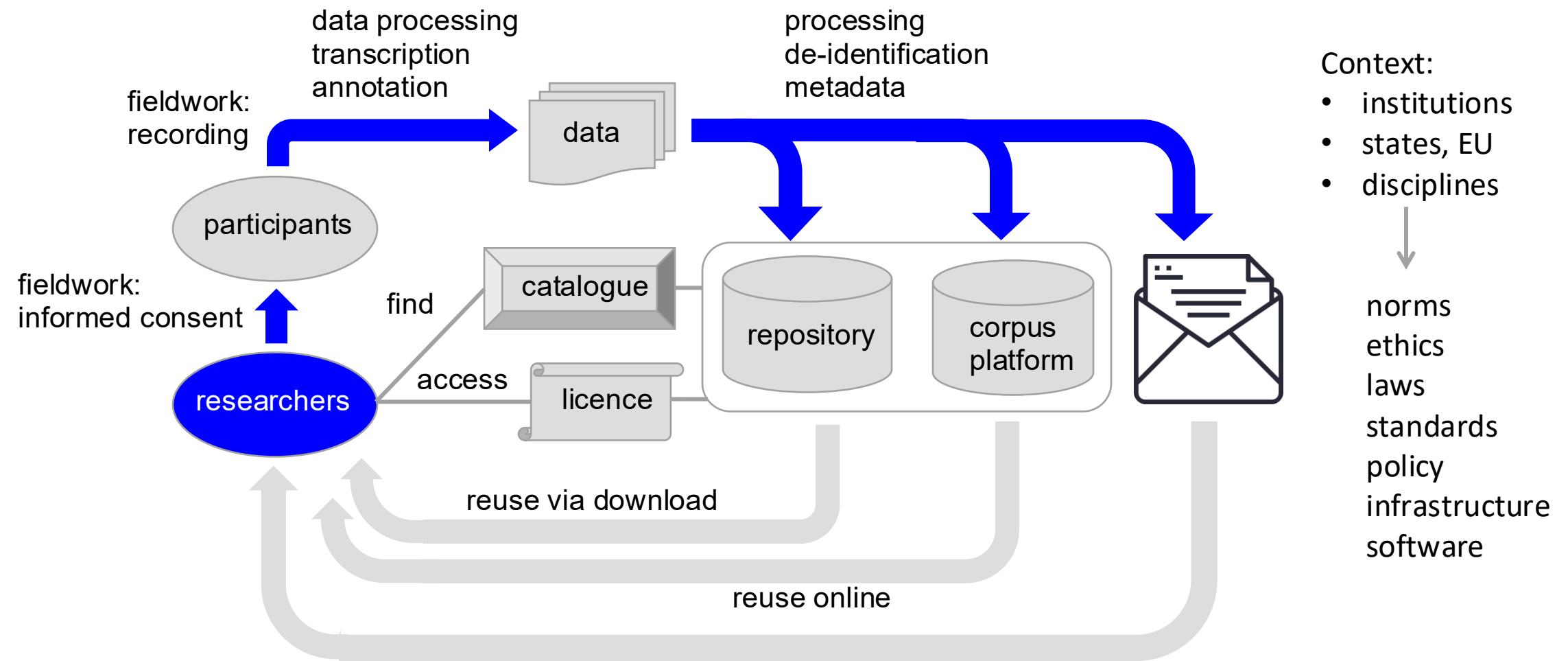


H. Hedeland & T. Schmidt 2022. The TEI-based ISO Standard ‘Transcription of spoken language’as an Exchange Format within CLARIN and beyond. *Selected papers from the CLARIN Annual Conference 2021*. Ed. M. Monachini & M. Eskevich. Linköping Electronic Conference Proceedings 189, pp. 34–45. DOI: <https://doi.org/10.3384/9789179294441>

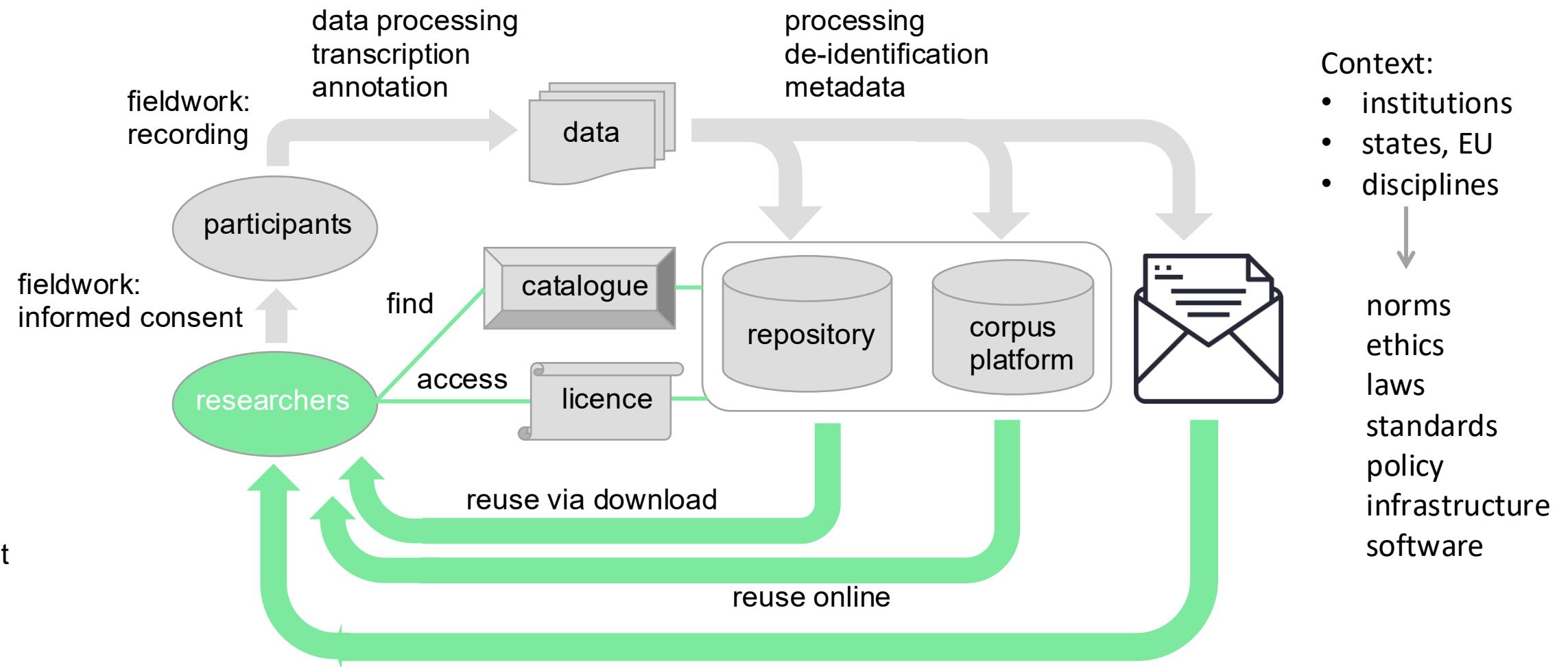
- Tokenizzazione:
  - divisione di un testo in “token”, cioè unità linguisticamente pertinenti di varia complessità
  - elaborazione indicata per l’inserimento in un database e ai fini della ricerca lessicometrica, morfo-sintattica, fonologica ecc.
  - input per procedure di annotazione automatica (p.es. l’annotazione delle parti del discorso PoS, la lemmatizzazione).
- Collaborazione in corso con *Linguisticbits* (Dr. Thomas Schmidt):
  - Ricongiungimento di frammenti di parola creati ai confini di segmenti in ELAN: procedura automatica basata su un lessico estratto da ItTenTen + correzione manuale
  - Concatenazione dei segmenti ELAN in ‘contributi’ (un’approssimazione dell’unità ‘turno’)
  - Divisione automatica dei contributi in parole
  - Conversione in un documento XML conforme al ISO/TC 37/SC 4, 2016 (standard ISO per la trascrizione della lingua parlata, codice CLARIN: application/tei+xml;format-variant=tei-iso-spoken;tokenized=[1]).

# 11. Condivisione dei dati

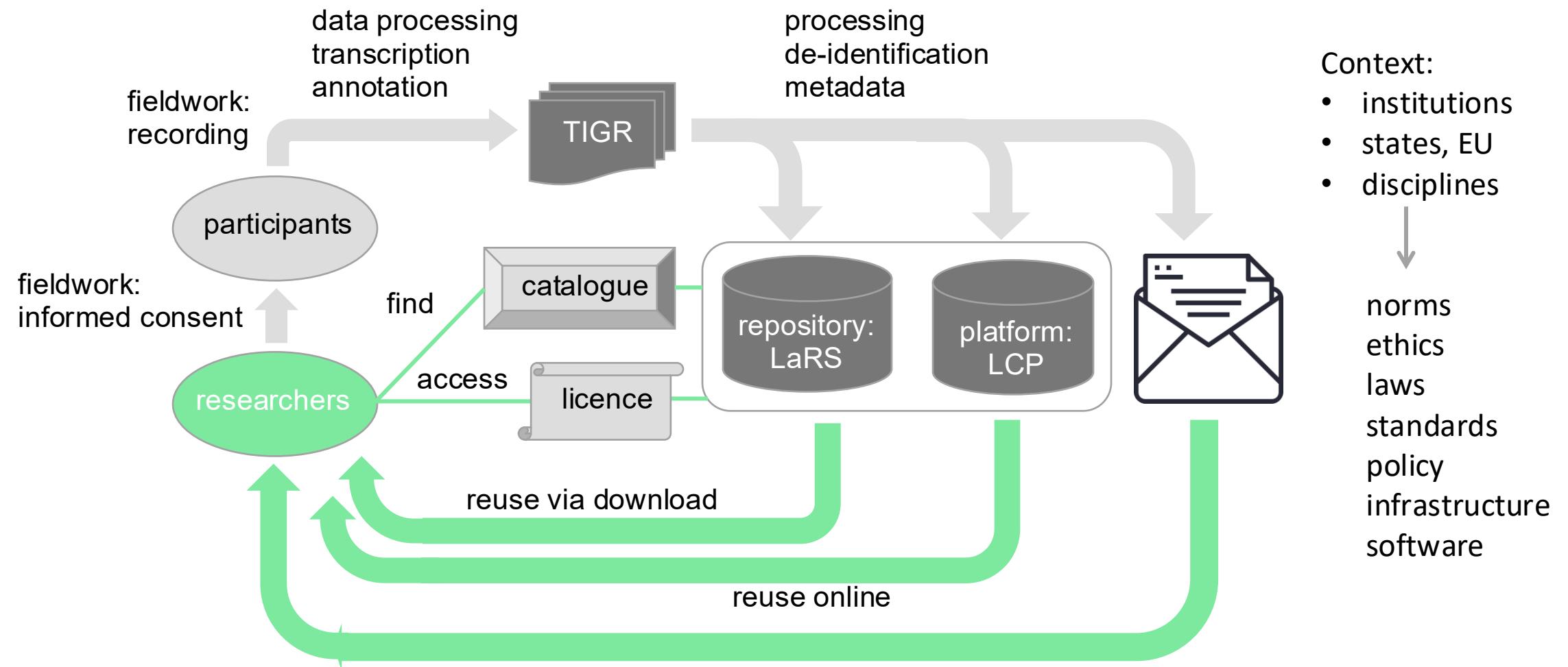
## Practices of data sharing and reuse



## Practices of data sharing and reuse



## Messa a disposizione e riuso del corpus TIGR



## Deposito dei dati sul repository LaRS @ SWISSUbase (<https://www.swissubase.ch/>)

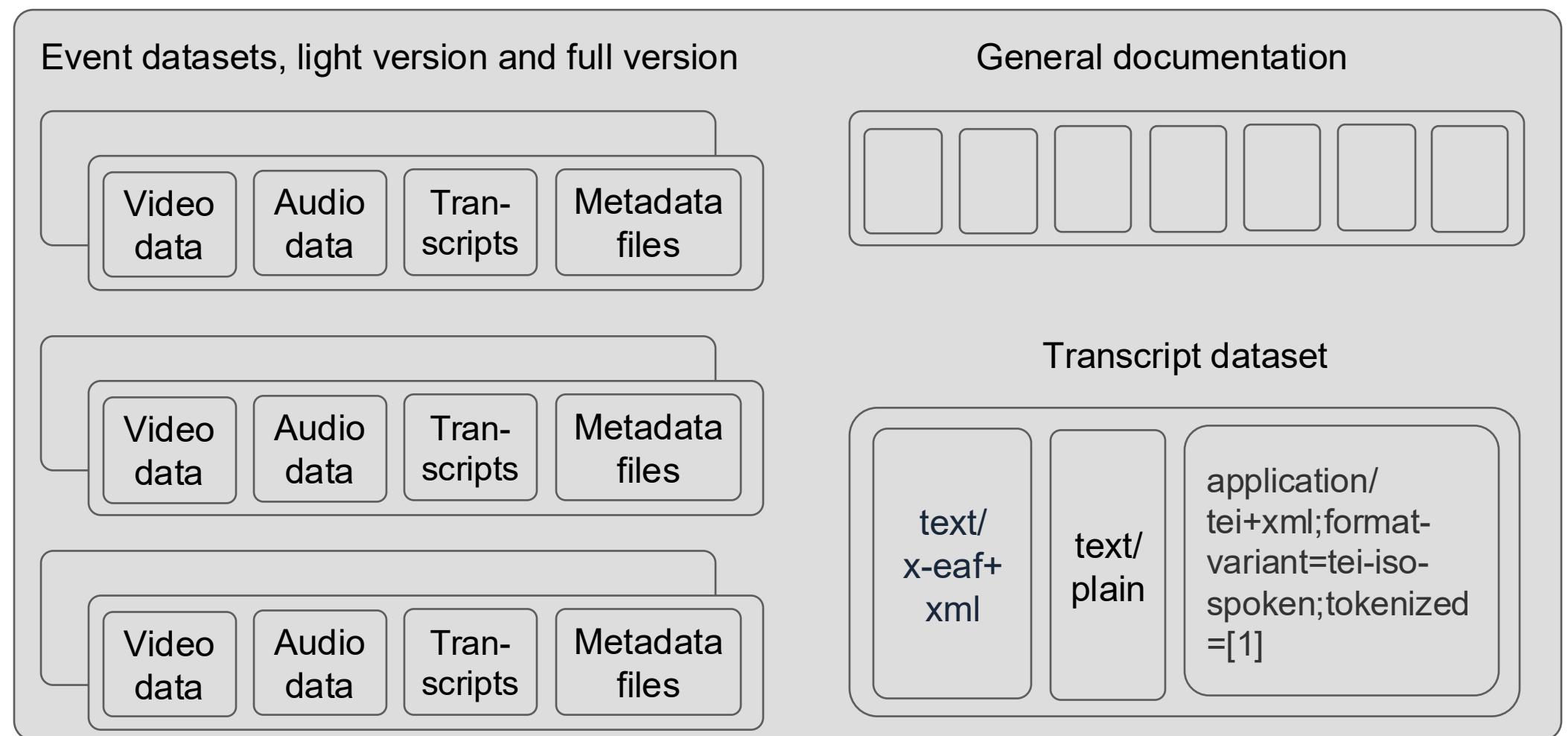
SWISSUbase.  
Metadata Guide for  
Linguistics Data.  
Metadata  
Documentation (last  
modified:  
21.11.2023).  
Language Repository  
of Switzerland  
(LaRS). [https://resources.swissubase.ch/linguistics\\_metadata-guide\\_en/](https://resources.swissubase.ch/linguistics_metadata-guide_en/)

Organizzazione a tre livelli:

- Studi
- Dataset (entità scaricabile)
- Documenti

Metadati obbligatori e facoltativi a ogni livello  
+ possibilità di inserire documenti appositi  
contenenti metadata

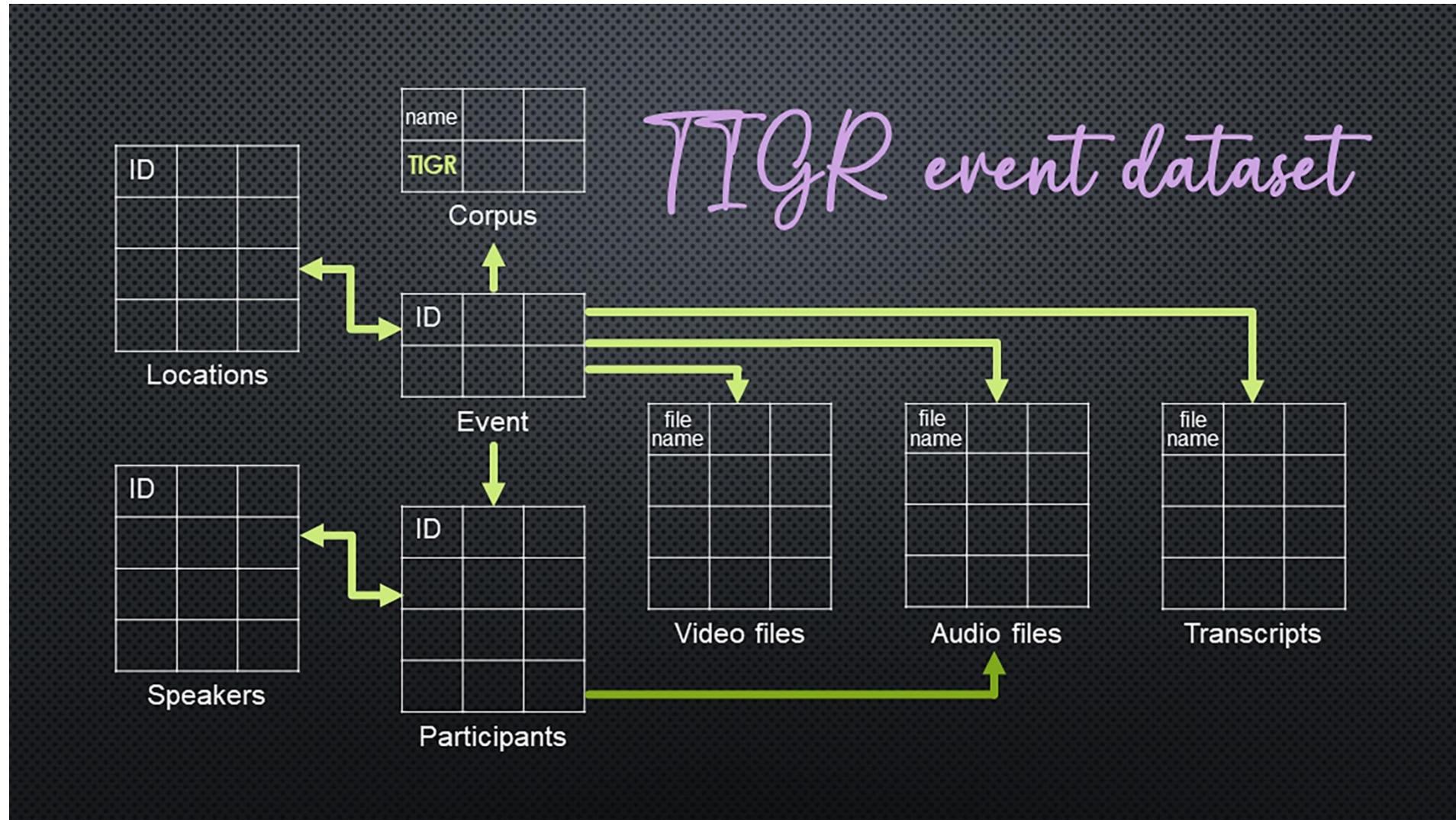
## TIGR on LaRS



## Documenti con metadati

Categorie ispirate  
da alcune  
componenti  
disponibili sul  
CMDI Component  
Registry (<https://catalog.clarin.eu/ds/ComponentRegistry#/>)

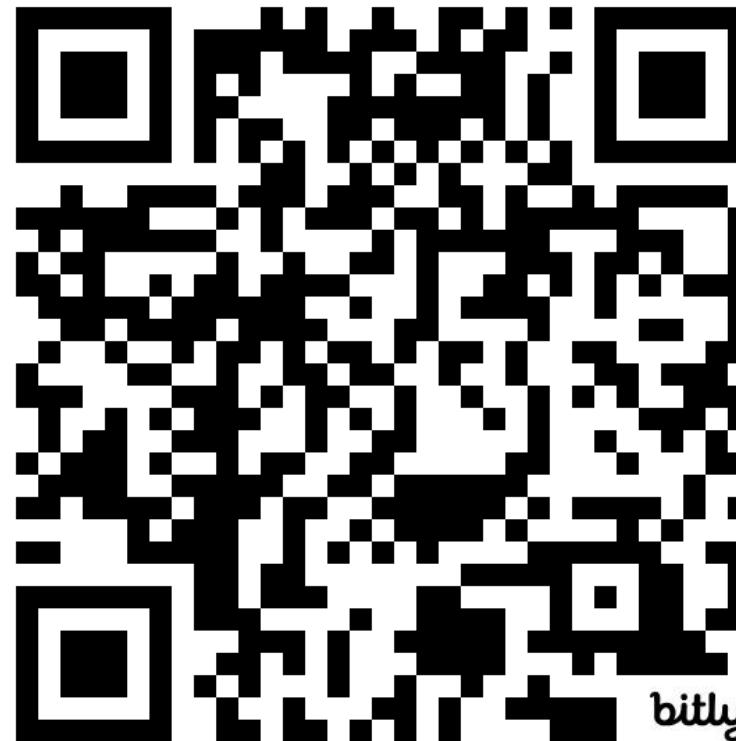
Cfr. anche i  
concetti CLARIN  
(Clarin Concept  
Registry CCR,  
<https://concepts.clarin.eu/CCR/>)



**Esempio: versione attuale dei dataset dell'evento 6a (link validi fino al 10 novembre 2024)**



ev-6a versione completa



ev-6a versione "leggera"

## Verso la messa a disposizione in una piattaforma digitale per la consultazione online

- Contesto istituzionale:
  - ShareTIGR
    - fra gli scopi: elaborazione delle trascrizioni TIGR in vista dell'inserimento in una piattaforma
  - ORD grant swissuniversities per FAIR-FI-LD, work package 2:  
collaborazione tra l'Università di Zurigo (Linguistic Research Infrastructure LiRI) e l'USI (Istituto di studi Italiani), luglio 2024-giugno 2025
    - scopi: installazione su un server USI di una istanza della Linguistic Corpus Platform LCP del LiRI, sviluppo dell'interfaccia di visualizzazione multimediale e di query *Videoscope* per adattatarla ai dati conversazionali
    - il corpus TIGR come caso studio
- In collaborazione con Linguisticbits: sviluppo di una procedura di upload in LCP di trascrizioni XML tokenizzate ISO/TC 37/SC 4, 2016.