ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

LABORATORIO
SPERIMENTALE

# Il corpus KIParla: nuovi formati e nuove prospettive di interoperabilità

**Silvia Ballarè, Caterina Mauri, Ludovica Pannitto & Eleonora Zucchini**

CLUB DAY – 19 maggio 2025

# INDICE


1. La costruzione del corpus KIParla: prassi
   - Difficoltà e problemi
   - Possibili (prime) soluzioni
2. Sviluppi e potenzialità
   - Un nuovo formato
   - Sistematizzazione e annotazione
3. Parlato e gestione FAIR dei dati
   - Soggettività e riproducibilità




ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

# INDICE


1. La costruzione del corpus KIParla: prassi
   - Difficoltà e problemi
   - Possibili (prime) soluzioni
2. Sviluppi e potenzialità
   - Un nuovo formato
   - Sistematizzazione e annotazione
3. Parlato e gestione FAIR dei dati
   - Soggettività e riproducibilità




ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

# La raccolta e la trascrizione dei dati: it takes a village!



A partire dal 2018, la raccolta e la trascrizione dei dati di parlato sono state affidate a studenti e studentesse che hanno preso parte al tirocinio KIParla (per un totale di **oltre 80 tirocinanti**).

Nel corso del tempo, è stata sviluppata una prassi che prevede:

- Formazione (per la raccolta e per la trascrizione);
- Incontri settimanali (in presenza o on-line) di monitoraggio;
- Contatti frequenti (via e-mail) per limitare disomogeneità e per coordinare la raccolta.

# Le trascrizioni

Le trascrizioni sono svolte manualmente: si tratta di parlato spontaneo, altamente interazionale con più lingue e più persone coinvolte.

Sono stati selezionati una serie di simboli (Jefferson 2004 + integrazioni) per dare conto di fenomeni tipici del parlato, come ad esempio:

| Symbol | Meaning |
|---|---|
| , | Slight rising intonation |
| ? | Sharp rising intonation |
| . | Final falling intonation |
| : | Prolonged vowel or consonant (one or two colons common, three or more colons only in extreme cases) |
| (.) | Micropause |
| = | End of one transcription unit and beginning of next begin with no gap/pause in between |
| °hello° | Syllables or words distinctly quieter than surrounding speech by the same speaker |
| HELLO | Syllables or words louder than surrounding speech by the same speaker |
| <hello> | Decreased speaking rate |
| >hello< | Increased speaking rate |
| [hello] | Speech overlaps |
| (hello) | Uncertain syllables or words |
| x | Inaudible syllables or words |
| # | Transcription unit containing at least one word in another language |
| ((laughs)) | Non-verbal behaviour |



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

# Problema numero 1: Disomogeneità/soggettività

La trascrizione *comporta* un primo livello di interpretazione del dato.

*Come si trascrive cè*? *Rusco* è italiano regionale o dialetto? Passa da una lingua all'altra ma come faccio a distinguerle? Il parlante di *scerto* anziché *certo*, come trascrivo? Qui devo interrompere l'unità?

- Regole generali (ad es. *trascuriamo ciò che è fonologico*, *trascriviamo ciò che è morfologico*)
  e discutiamo assieme i casi più problematici



- Accettazione di un margine di soggettività (v. allungamenti, rallentamenti e accelerazioni, …)

**Le trascrizioni del KIParla sono state effettuate da quasi 100 persone diverse (!).**

**Successivamente, ogni modulo è stato revisionato (e anonimizzato) da un'unica persona, in modo da uniformare (per quanto possibile) le disomogeneità e correggere gli errori.**



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

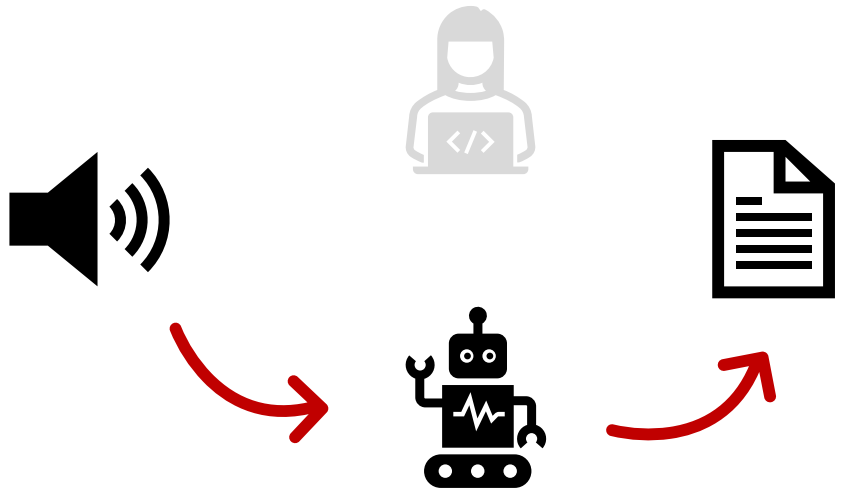
# Problema numero 2: Tempo

Le trascrizioni manuali richiedono **moltissimo** tempo.

(tradunt) Un trascrittore **esperto** impiega circa 10 ore a trascrivere un'ora di parlato.

A questo tempo, va sommato quello del trascrittore senior per la revisione.

Quindi abbiamo deciso di valutare la trascrizione automatica per verificare se potesse (o no) farci risparmiare del tempo.





ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

# Esperimento

Whisper – software di speech recognition di OpenAI:

- Livelli alti di performance
- Già sperimentato in precedenza
- Fornisce file in formato .srt (-> ELAN)

Setting sperimentale:

- 11 trascrittori volontari
- 10 minuti di audio manualmente + 10 minuti da trascrizione automatica
- 3 tipologie di conversazione
- 2 sessioni di lavoro da 2 ore

Protocollo di revisione trascrizione automatica:

- Minimo pre-processing dell'output asr
- Assegnazione parlanti e revisioni minime su txt
- Revisione finale su ELAN



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

# Esperimento



Analisi dei dati:

- Tempi: min. trascritti/30 min
- Confronto trascrizioni → **script**

Elaborazione di uno script per:

- Correggere automaticamente alcuni aspetti che non ci interessavano per l'analisi
- Estrarre le differenze delle due tipologie di trascrizione rispetto a una trascrizione 'gold'
- Organizzare le discrepanze in formato tabulare



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Domanda generale: l'introduzione di whisper nella pipeline influisce sulla capacità dei trascrittori di identificare gli aspetti prosodici? La "normalizzazione" imposta da whisper introduce un bias nelle trascrizioni?

# INDICE

1. La costruzione del corpus KIParla: prassi
   - Difficoltà e problemi
   - Possibili (prime) soluzioni

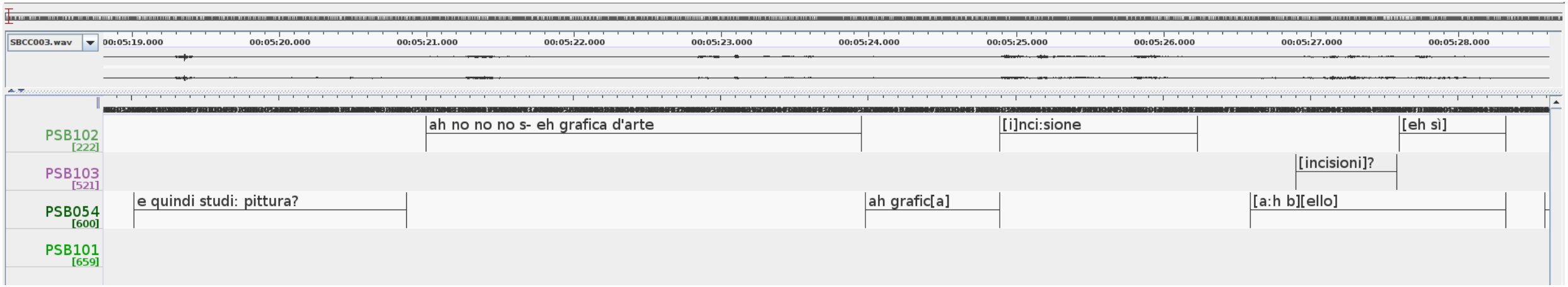
2. **Sviluppi e potenzialità**
   - **Un nuovo formato**
   - **Sistematizzazione e annotazione**

3. Parlato e gestione FAIR del dato
   - Soggettività e riproducibilità



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

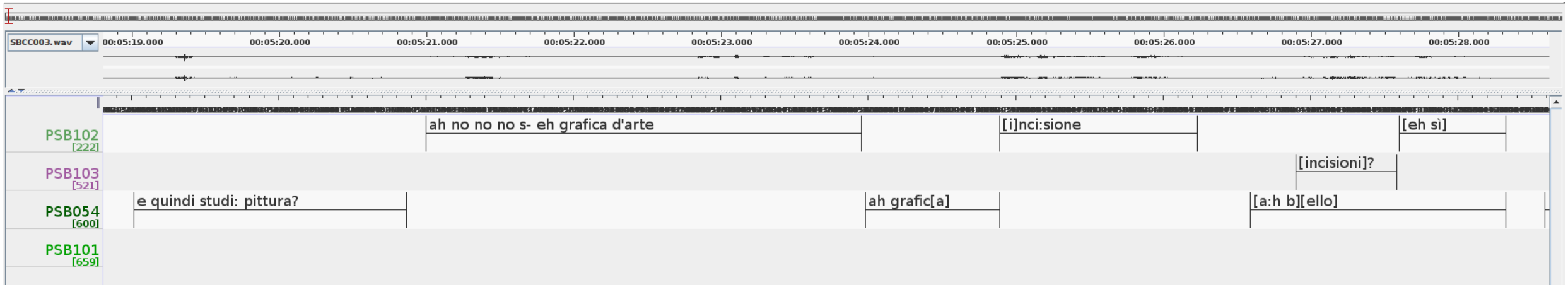
# - Unità di trascrizione allineate con l'audio






ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

- Unità di trascrizione allineate con l'audio
- Annotazione a livello di segmento

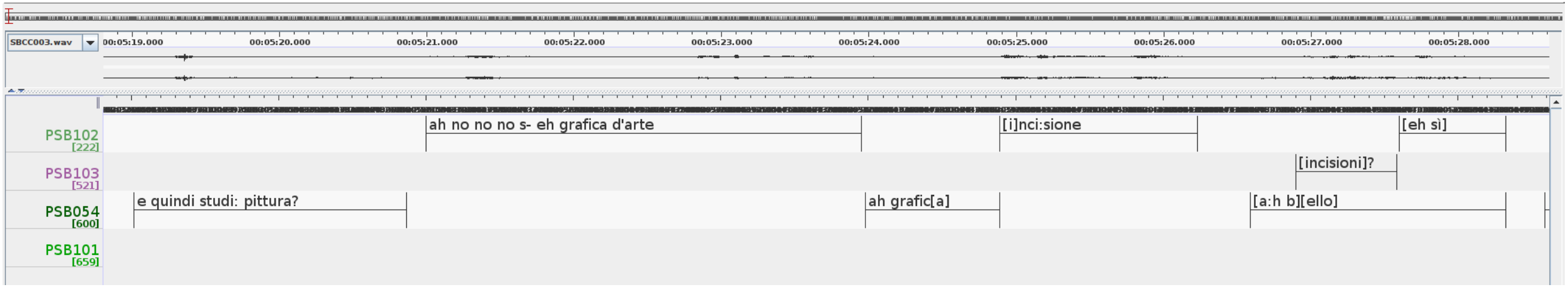



<lo:re:m [xx do::lor> sit am(et), °conse]cte[tur adipiscing] elit° sit amet consectetur? ((sigh)) adipisc~ ELIT quisque=faucibus ex adipiscing elit >'sque fa:ucibus ex sapien vi:tae pellentesque<



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

- Unità di trascrizione allineate con l'audio
- Annotazione a livello di segmento

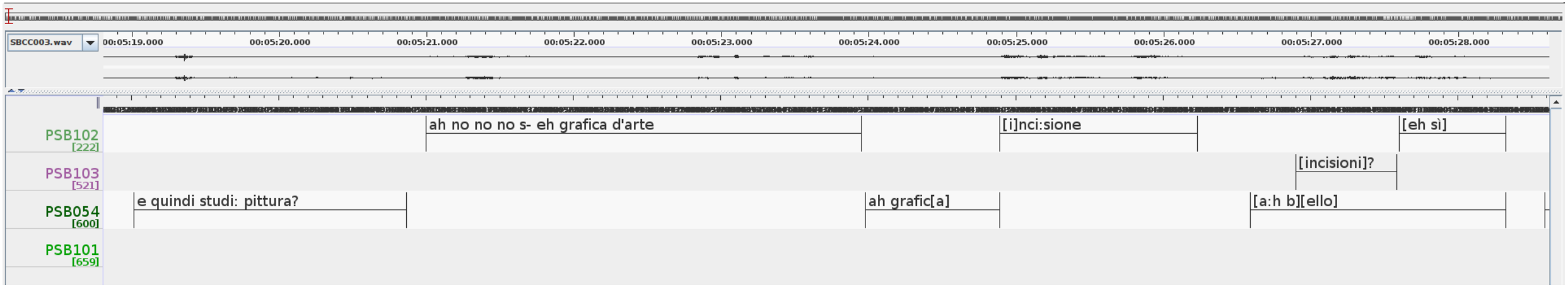



**<**lo:re:m **[xx** do::lor**>** sit am**(**et**),** °conse**]**cte**[**tur adipiscing**]** elit° sit amet consectetur**? ((sigh))** adipisc**~ ELIT** quisque**=**faucibus ex adipiscing elit **>'**sque fa:ucibus ex sapien vi:tae pellentesque**<**



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

- Unità di trascrizione allineate con l'audio
- Annotazione a livello di segmento
- Annotazione a livello di token

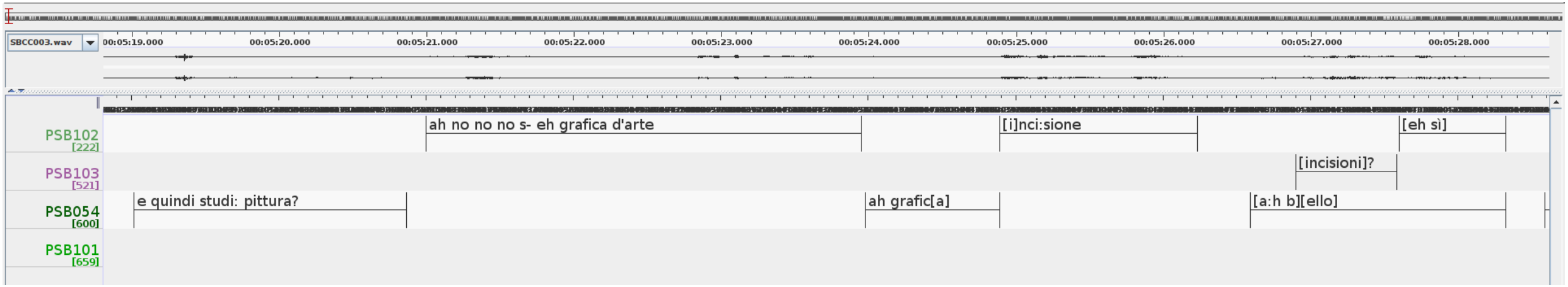



<lo:re:m [xx do::lor> sit am(et), °conse]cte[tur adipiscing] elit° sit amet consectetur? ((sigh)) adipisc~ ELIT quisque=faucibus ex adipiscing elit >'sque fa:ucibus ex sapien vi:tae pellentesque<



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

- Unità di trascrizione allineate con l'audio
- Annotazione a livello di segmento
- Annotazione a livello di token




Trascrizione con simboli conversazionali (Jefferson)

Trascrizione ortografica

Fortmato 'verticale' per NoSketch Engine



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

# Perché l’attuale pipeline non è ottimale

- Il processo richiede molto tempo e possono essere fatte solo ricerche per forme;
- Le trascrizioni sono realizzate a mano e non sono validate rispetto a un formato atteso;
- La preparazione dei dati è opaca (*se fossi un ricercatore, in quale formato vorresti questi dati?*);
- I dati derivati al momento non sono versionati;
- Non tutte le figure annotate sono *stabili* e aggiungere livelli di annotazione richiede nuovi formati.



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

# Trascrizione con simboli conversazionali (Jefferson) > CoNLL



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

# From Jefferson to CoNLL-compatible KIParla

1. Each transcription unit is **normalized** and **validated**

a. common errors and typos are corrected automatically (e.g., double spaces are reduced to one, number are spelled out…)

b. sequence level annotations are checked for consistency (e.g., parentheses balance)



19

# From Jefferson to CoNLL-compatible KIParla

2. Transcriptions units are **tokenized**

a. tokens are identified based on orthographic units
b. each token is assigned an id, a type (linguistic, unknown, non-linguistic-behavior, shortpause) and its boundaries in the transcription unit
c. linguistic tokens are checked against a regex (`['~-]?(\p{L}+:*)*\p{L}+:*[-'~]?[.,?]?`)
d. jefferson annotations are turned into token-level features



20

# From Jefferson to CoNLL-compatible KIParla

3. **Overlaps** among transcription units are computed and checked

a. Non-linguistic-behaviours are not considered as overlapping items

b. Overlapping events can involve more than two speakers

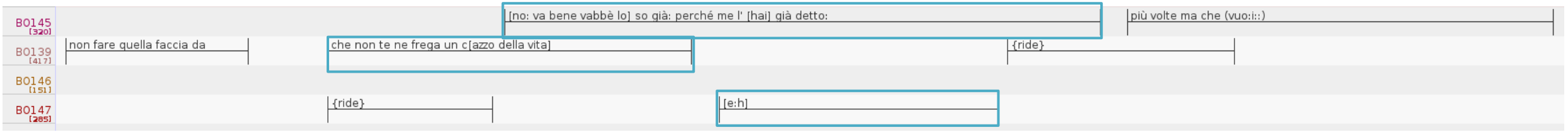
c. A transcription unit can participate in more than one overlapping event



21

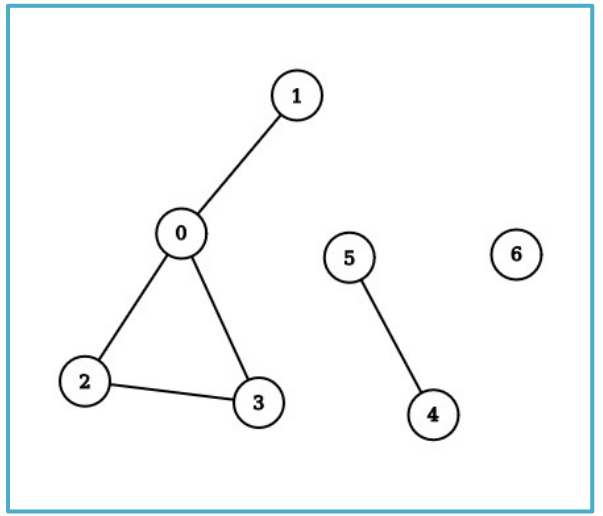
# Overlapping speech

## Case 1: unit participating in more than one overlapping event




## Case 2: overlap among three speakers






22

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

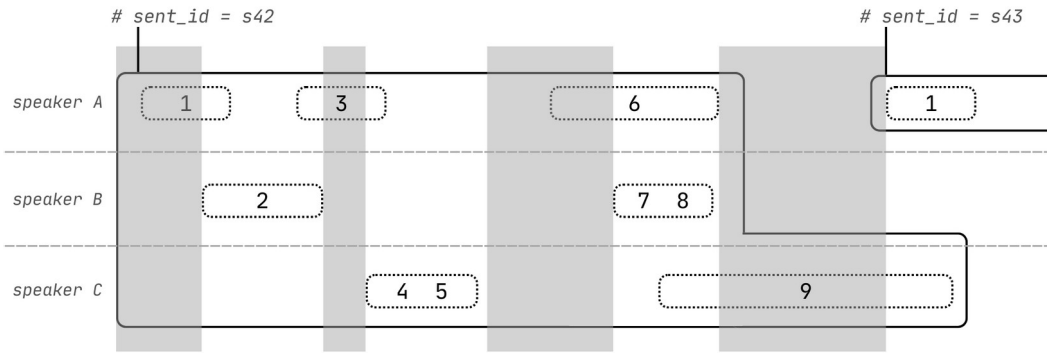
# (near-)Algorithm

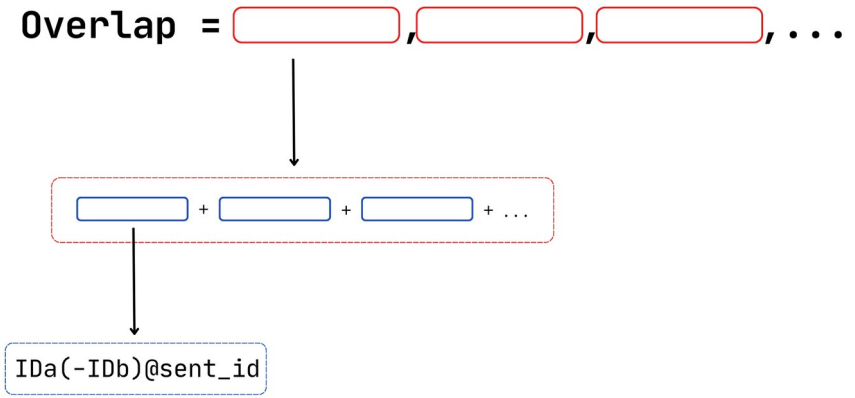
> Each unit is **linked** in a graph data structure to all units to which it overlaps

> Each transcription units also has a number of **character-spans** annotated as overlapping

> We identify **cliques** (i.e., strongly connected components) in the graph as overlapping events

> Overlapping events and overlapping spans are **matched**






23

# CoNLL-compatible KIParla: the fields

> Speaker ID

> Token ID (transcription unit ID + progressive ID)

> type (linguistic, metalinguistic, unknown, shortpause)

> original span

> CoNLL-U `sent_id`

> CoNLL-U fields

> alignment features

> jefferson-derived features



24

| | token_id | speaker | tu_id | unit | id | span | form | ... | type | jefferson_feats | align | prolong | pace | guesses | overlaps |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 38-0 | BO147 | 38 | _ | _ | (xx | x | _ | unknown | | Begin=49.327 | | _ | 0-2(0) | _ |
| 2 | 38-1 | BO147 | 38 | _ | _ | ma | ma | _ | linguistic | | _ | | _ | 0-2(0) | _ |
| 3 | 38-2 | BO147 | 38 | _ | _ | io | io | _ | linguistic | | _ | | _ | 0-2(0) | _ |
| 4 | 38-3 | BO147 | 38 | _ | _ | da | da | _ | linguistic | | _ | | _ | 0-2(0) | _ |
| 5 | 38-4 | BO147 | 38 | _ | _ | pallotti | pallotti | _ | linguistic | | _ | | _ | 0-8(0) | _ |
| 6 | 38-5 | BO147 | 38 | _ | _ | ci | ci | _ | linguistic | | _ | | _ | 0-2(0) | _ |
| 7 | 38-6 | BO147 | 38 | _ | _ | piglio | piglio | _ | linguistic | | _ | | _ | 0-6(0) | _ |
| 8 | 38-7 | BO147 | 38 | _ | _ | le | le | _ | linguistic | | _ | | _ | 0-2(0) | _ |
| 9 | 38-8 | BO147 | 38 | _ | _ | paste | paste | _ | linguistic | | _ | | _ | 0-5(0) | _ |
| 10 | 38-9 | BO147 | 38 | _ | _ | >cioè:< | cioè | _ | linguistic | | _ | 3x1 | Fast=0-4(0) | 0-4(0) | _ |
| 11 | 38-10 | BO147 | 38 | _ | _ | ci | ci | _ | linguistic | | _ | | _ | 0-2(0) | _ |
| 12 | 38-11 | BO147 | 38 | _ | _ | prend[o: | prendo | _ | linguistic | ProsodicLink=Yes | _ | 5x1 | _ | 0-6(0) | 5-6(39) |
| 13 | 38-12 | BO147 | 38 | _ | _ | mh | mh | _ | linguistic | | _ | | _ | 0-2(0) | 0-2(39) |
| 14 | 38-13 | BO147 | 38 | _ | _ | le | le | _ | linguistic | | _ | | _ | 0-2(0) | 0-2(39) |
| 15 | 38-14 | BO147 | 38 | _ | _ | brio~]) | brio~ | _ | linguistic | Interrupted=Yes | End=52.997 | | _ | 0-5(0) | 0-5(39) |
| 16 | 39-0 | BO146 | 39 | _ | _ | [le | le | _ | linguistic | | Begin=51.927 | | _ | _ | 0-2(38) |
| 17 | 39-1 | BO146 | 39 | _ | _ | lasagne] | lasagne | _ | linguistic | | End=52.887 | | _ | _ | 0-7(38) |



25

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

# APPUNTI

1. Stato attuale e limiti (lavoro manuale, → SILVIA
   miglioramenti del processo (work in progress):

- esperimento per automatizzazione trascrizione → ELE ok
- **script uniformazioni** → LUDO Già FATTO

Limiti:

Tempo per trascrizione

Tempo per correzione

Disomogeneità (tanti trascrittori, anche se un revisore per modulo e condivisione pratiche)

Quindi abbiamo fatto esperimento per omogeneità e risparmio tempo

Non abbiamo i risultati MA per confrontare trascrizioni manuali e automatizzate, abbiamo sviluppato script per pescare errori manuali comuni → questo in realtà ci permette uniformazione e poi nuovi formati



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

**APPUNTI**

2. Sviluppi e potenzialità (work in progress)
Nuovo formato

→ conteggio automatizzato di cose, correzione sistematica

→ LUDO

→ **diversi livelli di annotazione (ad es. pos, lemm e UD)**

→ LUDO Già FATTO

**3.** problemi FAIR del parlato(riproducibilità) / soggettività → CATE



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

# Cosa possiamo fare adesso

- Agreement
- Treebank
- Lemmatizzazione
- Rollback
- Versioning
- Grafo di dipendenze tra file



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

# INDICE


1. La costruzione del corpus KIParla: prassi
   - Difficoltà e problemi
   - Possibili (prime) soluzioni

2. Sviluppi e potenzialità
   - Un nuovo formato
   - Sistematizzazione e annotazione

**3. Parlato e gestione FAIR dei dati**
   - **Soggettività e riproducibilità**




ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

# Dalla trascrizione all'analisi: cosa ci dicono le differenze?

- Le trascrizioni manuali presentano variabilità significative: ogni trascrittore ha un certo margine di soggettività su confini delle unità, proprietà prosodiche, categorie di annotazione.

- Gli script sviluppati possono rivelare *zone sistematicamente instabili*, che dipendono sia dal tipo di parlato sia dal tipo di annotazione, e zone più omogenee.

- Invece di rimuovere questa variabilità, la possiamo rendere *osservabile*



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

# Rappresentare la soggettività, in modo trasparente

- I nuovi formati multilivello permettono di mantenere tracciabilità completa: ogni modifica è registrata, ogni trascrittore e annotatore identificabile (Git)
- La gestione di versioni parallele consente di documentare il disaccordo e di confrontare approcci annotativi diversi.
- Si passa da una visione "normativa" dell'annotazione a una visione dialogica e pluralista.



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

# Riproducibilità e apertura: cosa cambia

- L’adozione di protocolli documentati e versionabili rende il processo di annotazione riproducibile e reversibile.
- I dati così organizzati sono più interoperabili, accessibili e riutilizzabili anche da chi non ha partecipato alla loro costruzione.
- Questo approccio supporta analisi quantitative su instabilità categoriali, confronti tra annotatori, e sviluppo di metriche per la soggettività.



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

# Applicare i principi FAIR al parlato

- Findable: Utilizzo di metadati ricchi e identificatori persistenti per rendere i dati parlati facilmente rintracciabili.
- Accessible: Definizione chiara delle condizioni di accesso, considerando le implicazioni etiche e legali.
- Interoperable: Adozione di formati standard e vocabolari condivisi per facilitare l'integrazione con altri dati e strumenti.
- Reusable: Documentazione dettagliata e licenze chiare per permettere il riutilizzo dei dati in diversi contesti.



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

# Verso una rappresentazione responsabile del parlato

- Implementare protocolli che garantiscano la tracciabilità delle modifiche e delle decisioni annotative.
- Promuovere la trasparenza nelle scelte metodologiche per rafforzare la fiducia nella qualità dei dati.
- Favorire la collaborazione tra ricercatori per sviluppare linee guida condivise e adattabili alle diverse esigenze.



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA




**Grazie! Domande?**



