ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

LABORATORIO
SPERIMENTALE

Il corpus KIParla: nuovi formati e nuove prospettive di interoperabilità

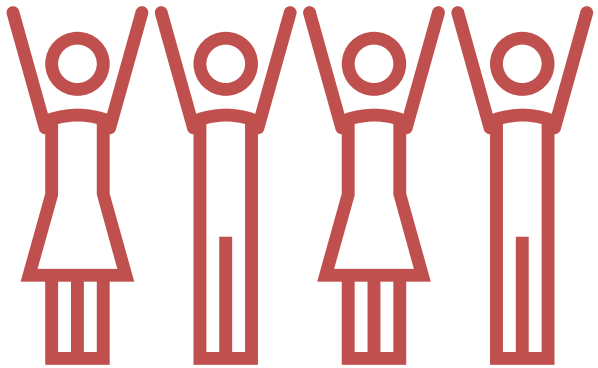
Silvia Ballarè, Caterina Mauri, Ludovica Pannitto & Eleonora Zucchini

CLUB DAY – 19 maggio 2025

INDICE

INDICE

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.

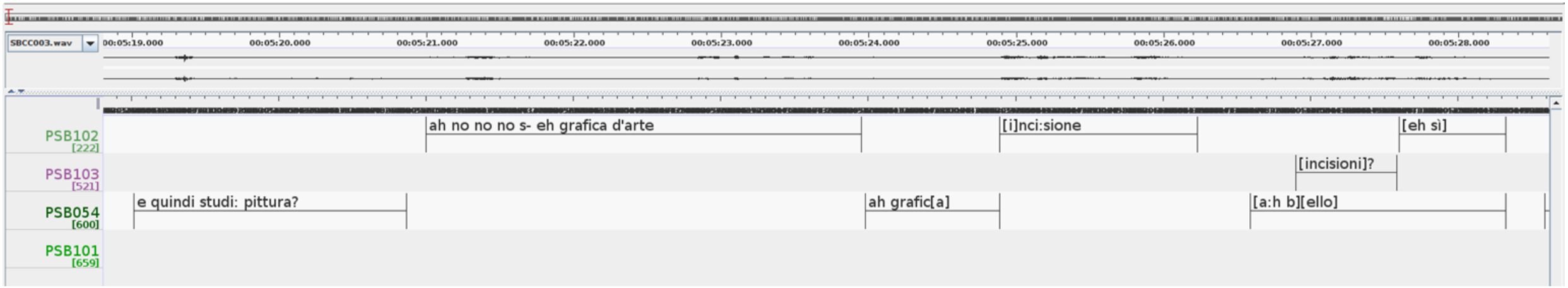
Le trascrizioni

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
$	non-existent word in Italian
((laughs))	Non-verbal behaviour

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

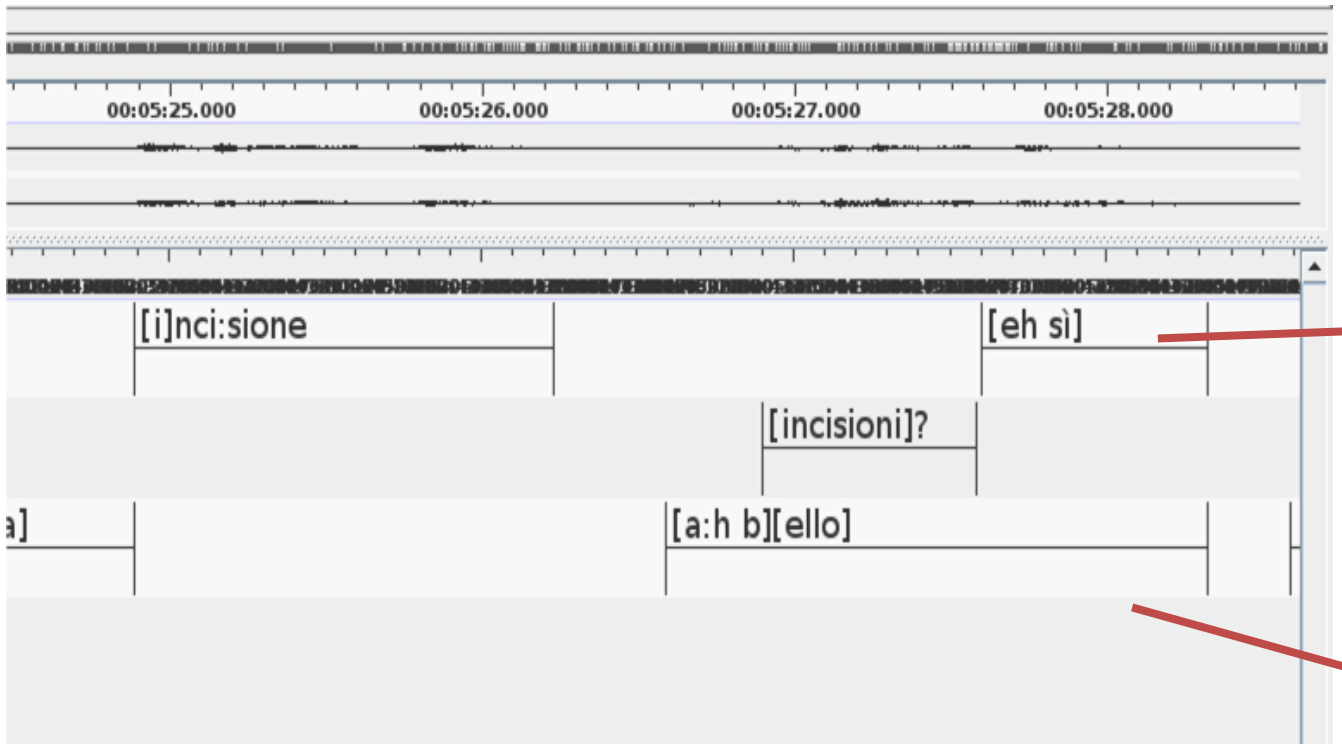
Sono stati selezionati una serie di simboli (Jefferson 2004 + integrazioni) per dare conto di fenomeni tipici del parlato

1. Unità di trascrizione allineate con l'audio



1. Unità di trascrizione allineate con l'audio

2. Feature a livello di span



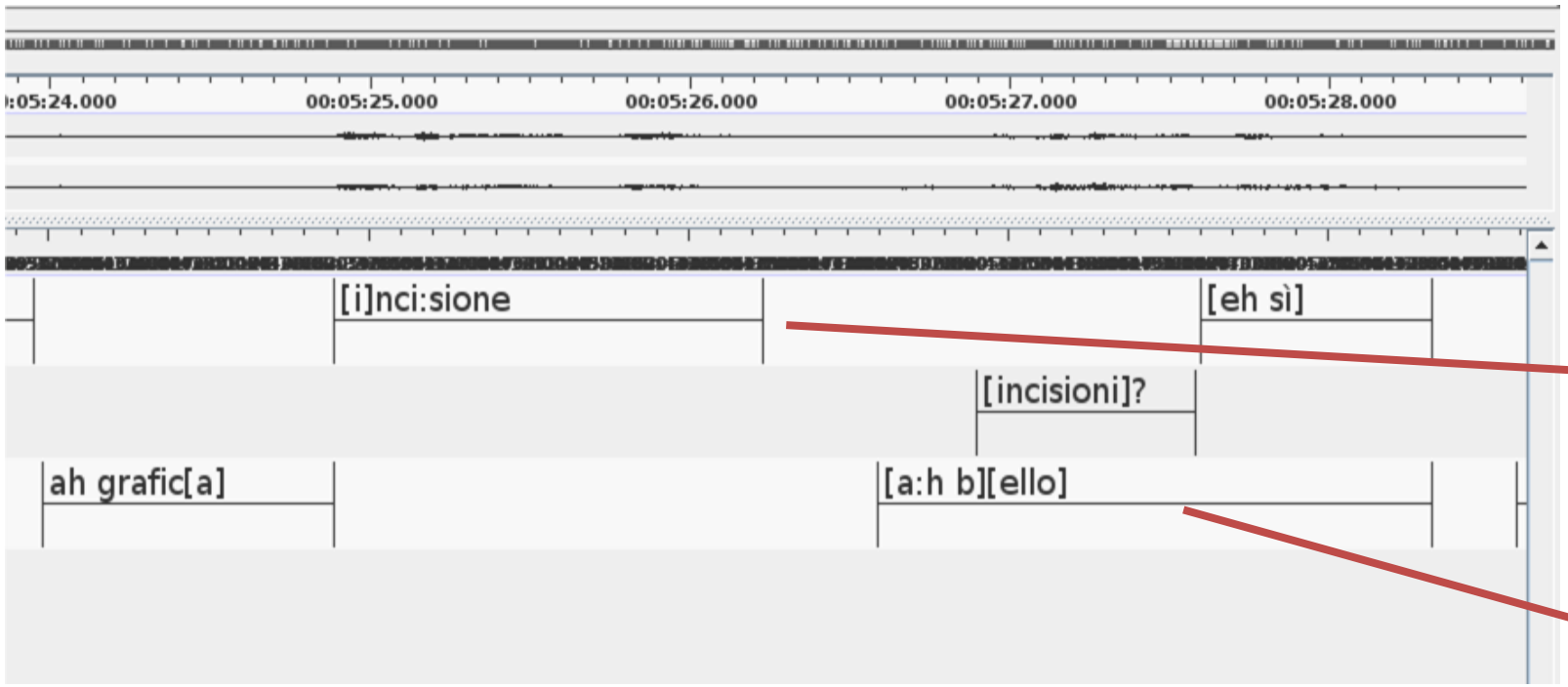
[eh sì]

[a:h b][ello]

1. Unità di trascrizione allineate con l'audio
2. Feature a livello di span
3. **Feature a livello di token**

```
xxx poi
```

```
((ride))
```



[i]nci:sione

[a:h b][ello]

Problema numero 1: Tempo

Le trascrizioni manuali richiedono **moltissimo** tempo.

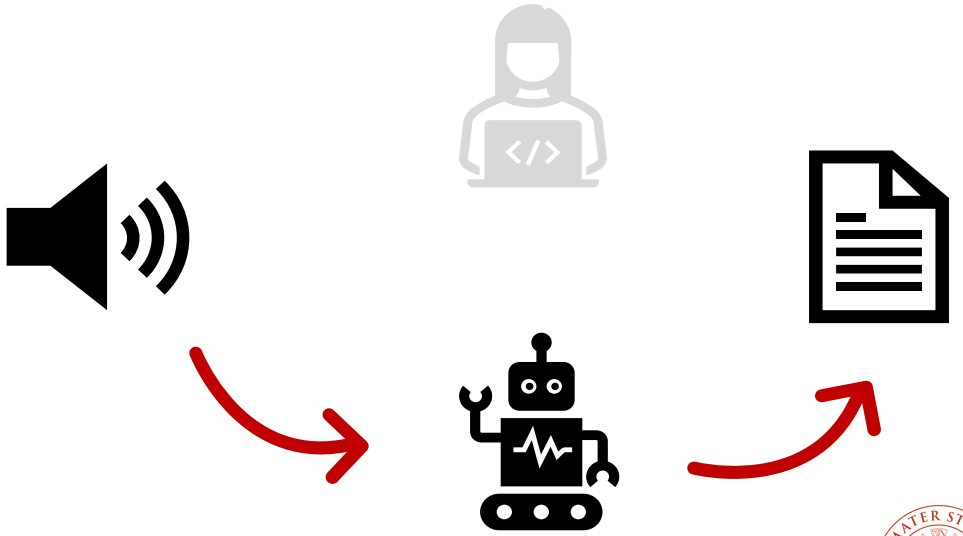
T > 10:1

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

(tradunt)

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

DOMANDA:
La trascrizione automatica può farci risparmiare del tempo?



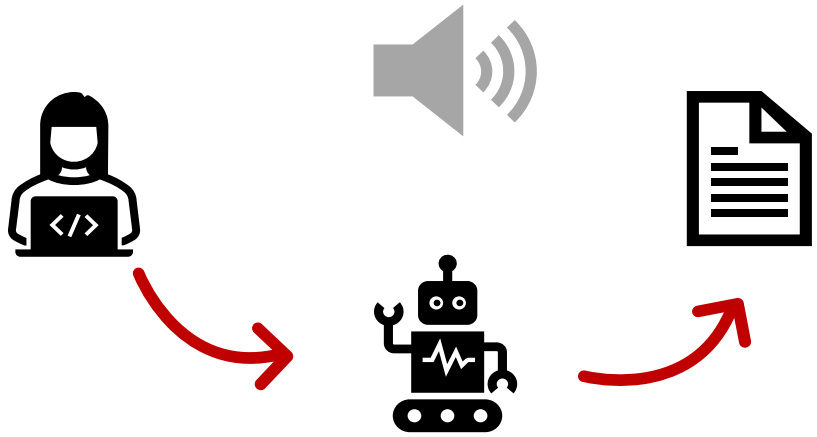
Problema numero 2: Disomogeneità/soggettività

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?

DOMANDA:
Possiamo migliorare la coerenza interna al dato?

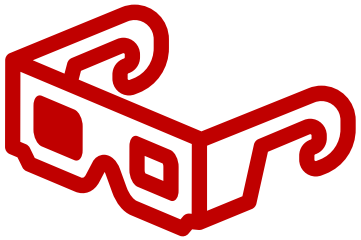
- Forniamo **regole generali** (ad es. *trascuriamo ciò che è fonologico, trascriviamo ciò che è morfologico*) e discutiamo assieme i casi più problematici
- Accettiamo la presenza di un **margine di soggettività** (v. allungamenti, rallentamenti e accelerazioni, …)



Ad oggi hanno trascritto **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.

Problema numero 3: Formati e versioni



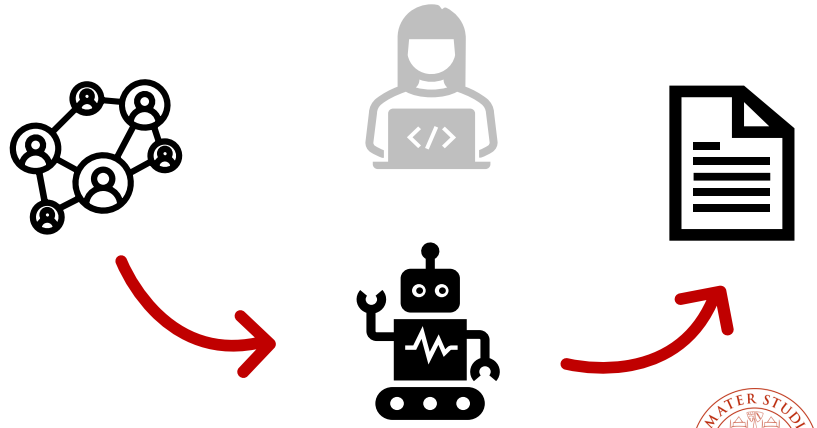
Ci sono già tantissime informazioni a cui si può accedere (allineamento, features conversazionali, trascrizione...), ma è difficile incrociare informazioni provenienti da query diverse.

DOMANDA:
Possiamo migliorare l'interoperabilità tra le varie modalità di accesso al corpus, e aggiungerne di nuove?

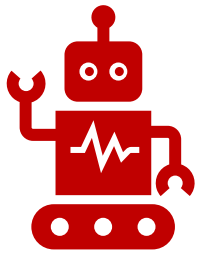
1. La trasformazione dei dati da un formato all'altro e la relazione che c'è tra questi è opaca all'utente (ad es. .eaf > vert)

2. I dati al momento non sono versionati;

3. Ogni nuovo livello di annotazione richiede un nuovo formato o una modifica sostanziale alla pipeline attuale.



Problema 1 → Esperimento



Software di *automatic speech recognition* (ASR) di OpenAI:

- Livelli alti di performance
- Utilizzabile in locale senza necessità di trasferire dati a servizi esterni
- Fornisce file in formato .srt (-> ELAN)

<u>Setting sperimentale:</u>

- 11 trascrittori volontari
- 2 sessioni di lavoro da 2 ore
- **Fase 1:** 10 minuti di audio trascritto manualmente
- **Fase 2:** 10 minuti di audio da revisionare dopo la trascrizione automatica

<u>Protocollo di revisione trascrizione automatica:</u>

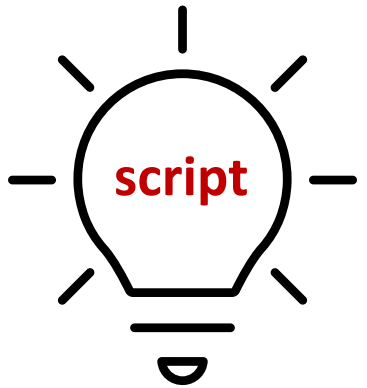
- Minimo pre-processing dell'output ASR
- Assegnazione parlanti e revisioni minime su file testuale
- Creazione di 1 .srt / parlante
- Revisione finale su ELAN

Esperimento: analisi

L'introduzione di *whisper* nella pipeline influisce sulla capacità dei trascrittori di identificare gli aspetti prosodici? La standardizzazione prodotta da whisper introduce un bias nelle trascrizioni?

Analisi dei dati:

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



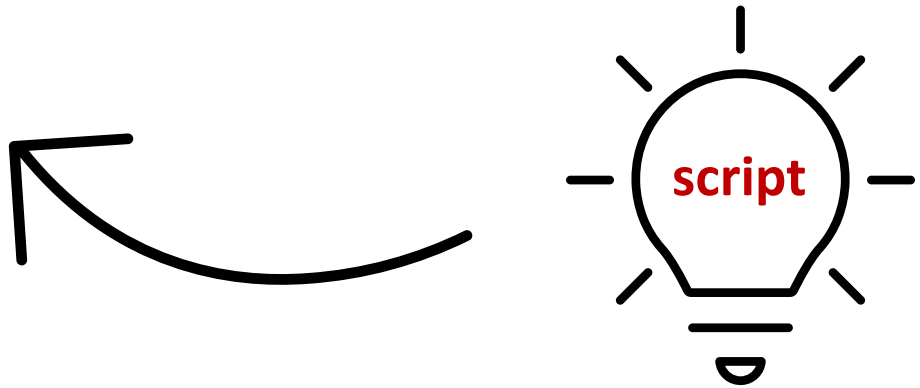
Elaborazione di uno script per:

- 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

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

L'introduzione di whisper cambia qualcosa per...

1. Il numero di minuti trascritti in due ore
2. Il numero di token linguistici trascritti, il numero di pause brevi, il numero di comportamenti metalinguistici
3. Il numero di unità di trascrizione, la loro composizione (e.g., unità che contengono solo comportamenti metalinguistici) e la loro durata in termini di token e di millisecondi
4. Il numero di overlap individuati
5. Il numero di elementi prosodici annotati
6. Attribution delle unità di trascrizione ai parlanti
7. L'identificazione delle unità di backchannel
8. Il contenuto della trascrizione:

- Se c'è un MISMATCH tra Fase 1 e Fase 2, c'è una categoria di token su cui l'introduzione di whisper pesa di più? Le discrepanze sono dovute a typos, standardizzazioni introdotte da whisper o altro...? C'è una posizione in cui il mismatch è più probabile (e.g., al confine dell'unità di trascrizione)?
- Se sia nella fase 1 che nella fase 2 è stato trascritto lo stesso token, è stato anche annotato nello stesso modo a livello prosodico? Ci sono features che sono più predittive di altre, sulla possibilità che esistano mismatch in altre colonne? (e.g., *una bassa accuratezza sui prolungamenti è predittiva rispetto all'accuratezza sul volume della voce?*)

9. Inoltre whisper fornisce la probabilità di predizione di ogni token: c'è una correlazione tra l'intervento del trascrittore/revisore e la probabilità di predizione?

La pipeline attuale

13 TOD2016 a sono più vissuta da sola // è più grande chiaramente // eh okay **perfetto** // quindi come dire // non mi pare che tu abbia una preferenza o

14 PTB019 // okay abbiamo finito grazie mille // mh prego // gentilissima // **perfetto** // allora // lei non è di torino // no // no // è nato // a foggia // roc

15 PBA004 anquilla tranquilla noi siamo due xxx ci ci difendiamo xxxxxxxx // **perfetto** // ehm // no ecco poi molti posti // quando vai nelle città // quind

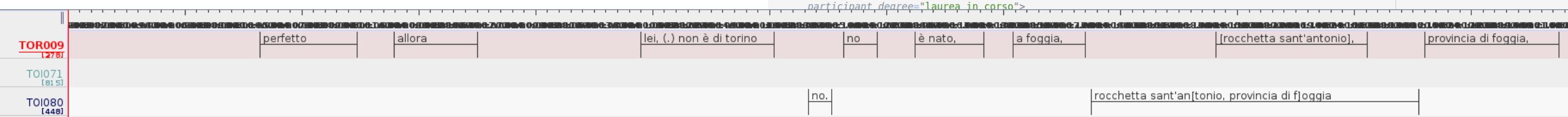
in teoria dovrebbero venire degli amici da roma // una mia cugina da ferrara // e // un nostro amico qua dal piemonte // e dovremmo festeggiare lì // a casa di un mio amico // va bene // okay abbiamo finito grazie mille // mh prego // gentilissima // **perfetto** // allora // lei non è di torino // no // no // è nato // a foggia // rocchetta sant'antonio provincia di foggia // rocchetta sant'antonio // provincia di foggia // quando si è traferito a torino // ah nel // cinquantotto // e quanti anni aveva // avevo

```
[kiparla][vps-a25197b6:vert] $ ls -lh
total 276M
-rw-rw-r-- 1 kiparla kiparla  33M Apr 15 09:48 KIP
-rw-rw-r-- 1 kiparla kiparla 134M Apr 15 09:49 KIPARLA
-rw-rw-r-- 1 kiparla kiparla  33M Apr 15 09:48 KIPasti
-rw-rw-r-- 1 kiparla kiparla  41M Apr 15 09:48 ParlaBO
-rw-rw-r-- 1 kiparla kiparla 2.8M Apr 15 09:49 ParlaBZ
-rw-rw-r-- 1 kiparla kiparla  34M Apr 15 09:49 ParlaTO
```

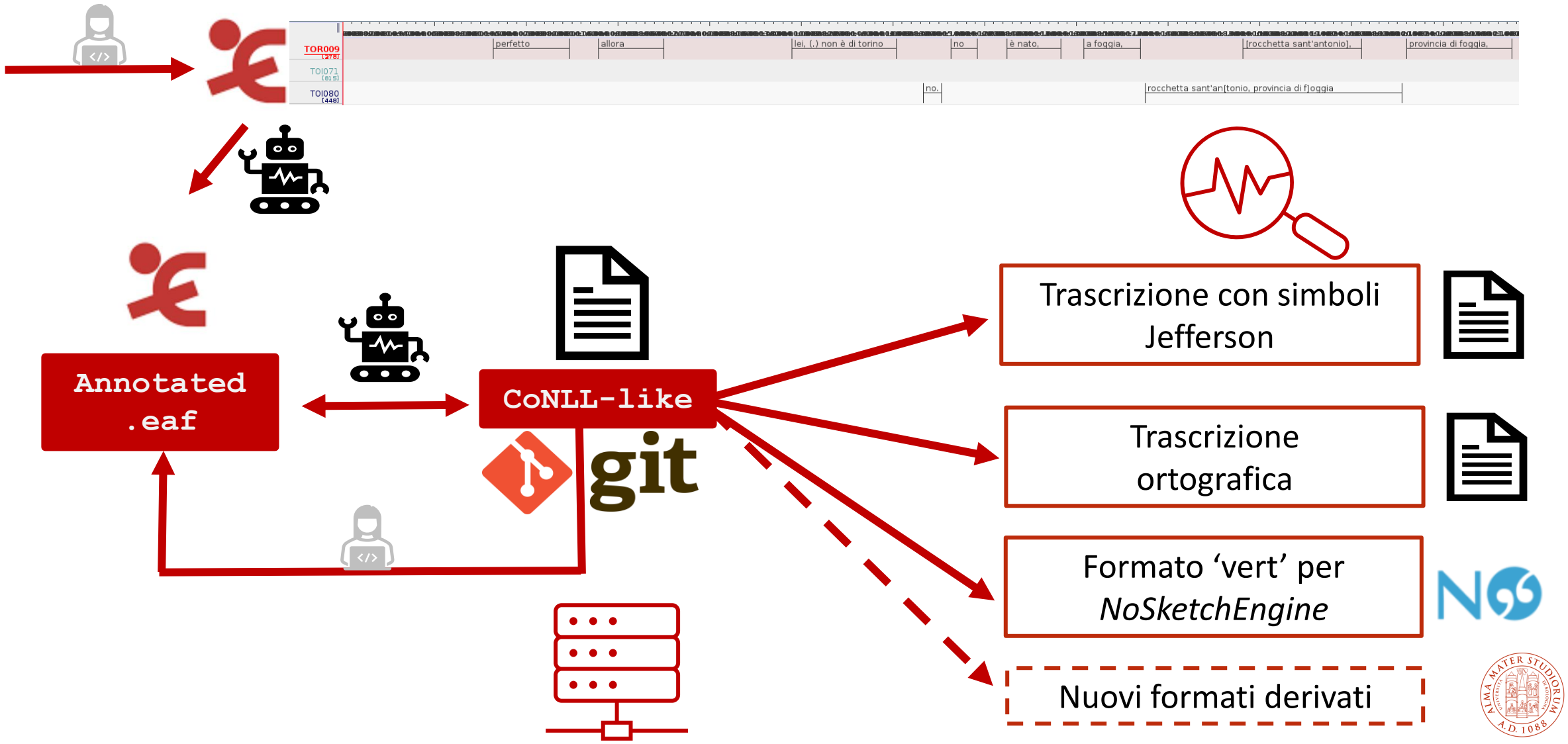
```
Conversation: PTB019

TOR009  perfetto
TOR009  allora
TOR009  lei non è di torino

TOI080  no

TOR009  no
TOR009  è nato
TOR009  a foggia

TOI080  rocchetta sant'antonio provincia di foggia
```

```
Conversation: PTB019

TOR009  perfetto
TOR009  allora
TOR009  lei, (.) non è di torino

TOI080  no.

TOR009  no
TOR009  è nato,
TOR009  a foggia,

TOI080  rocchetta sant'an[tonio, provincia di f]oggia
```

```
grazie
mari
//
</annotation>
<annotation begin="1820537" end="1821097" audio_file="https://search.corpuskiparla.it/corpus/player/player.cgi?code=TOD2007&begin=1820537&end=1821097" participant_code="TO043" participant_occupation="stud" participant_sex="F" files_in_which_participant_appears="TOD2007" participant_birth_region="piemonte" participant_age_range="21-25" participant_degree="laurea in corso">
sì
cia~
//
</annotation>
<doc doc_number="PTB019" full_conversation="https://search.corpuskiparla.it/corpus/conversations/PTB019.html" full_jefferson="https://search.corpuskiparla.it/corpus/conversations/PTB019_jefferson.html">
<conversation type="intervista semistrutturata" code="PTB019" duration="00:43:39" participants_number="3" languages="italiano-dialetto" participants="TOI071,,TOI080,,TOR009">
<annotation begin="1640" end="2470" audio_file="https://search.corpuskiparla.it/corpus/player/player.cgi?code=PTB019&begin=1640&end=2470" participant_code="TOR009" participant_occupation="stud" participant_sex="F" files_in_which_participant_appears="PTB019" participant_birth_region="sicilia" participant_age_range="21-25" participant_degree="laurea in corso">
perfetto
//
</annotation>
<annotation begin="2790" end="3500" audio_file="https://search.corpuskiparla.it/corpus/player/player.cgi?code=PTB019&begin=2790&end=3500" participant_code="TOR009" participant_occupation="stud" participant_sex="F" files_in_which_participant_appears="PTB019" participant_birth_region="sicilia" participant_age_range="21-25" participant_degree="laurea in corso">
allora
//
</annotation>
<annotation begin="4900" end="6040" audio_file="https://search.corpuskiparla.it/corpus/player/player.cgi?code=PTB019&begin=4900&end=6040" participant_code="TOR009" participant_occupation="stud" participant_sex="F" files_in_which_participant_appears="PTB019" participant_birth_region="sicilia" participant_age_range="21-25" participant_degree="laurea in corso">
```



La pipeline futura



Trascrizione con simboli conversazionali > CoNLL

Ogni unità di trascrizione viene normalizzata e validata

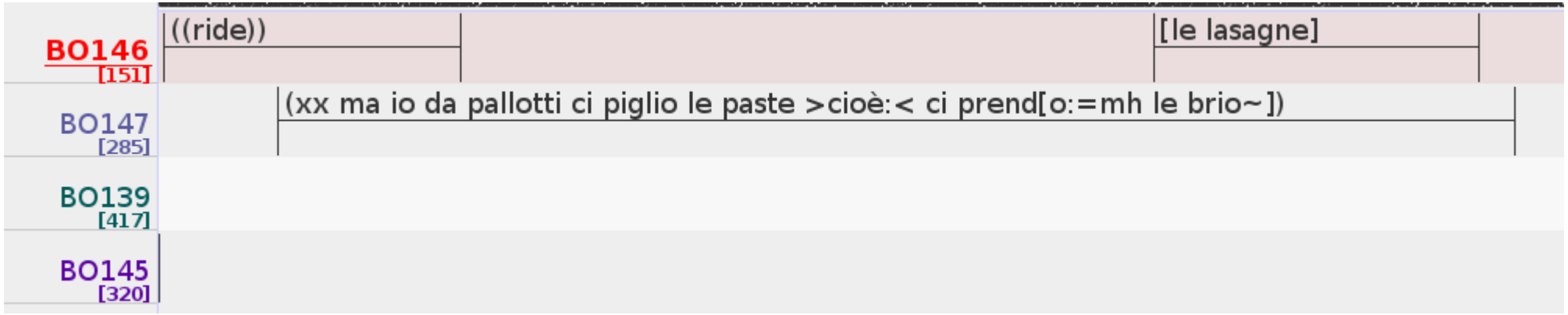
- Gli errori comuni sono corretti automaticamente (ad es. doppi spazi, parentesi non chiuse, numeri, …);
- Viene controllata la conformità della trascrizione rispetto al formato dichiarato (ad es. Bilanciamento delle feature a livello di span).

Le unità di trascrizione vengono tokenizzate

- I token (i.e., unità base) sono identificati secondo le unità ortografiche;
- Ad ogni token è assegnato un ID, un tipo (*linguistico*, *irriconoscibile*, *comportamento-non-linguistico*, *pausa*) e lo span di caratteri corrispondente nell'unità di trascrizione;
- I token linguistici sono validati con un'espressione regolare configurabile
- Le annotazioni conversazionali sono convertite in features a livello di token

Gestione e controllo delle sovrapposizioni

- Il formato Jefferson è sottospecificato rispetto alla presenza di sovrapposizioni > Induciamo i livelli di informazione impliciti.



	TID	SPEAKER	SPAN	FORM	TYPE	TOKEN_FEATS	ALIGNMENT	SPAN_FEATURES	OVERLAP	OVERLAP_SPAN
2	37-0	BO146	{ride}	{ride}	metalinguistic		End=49.872	_	_	_
3	38-0	BO147	(xx	x	unknown		Begin=49.327	6:Guess=0-2	_	_
4	38-1	BO147	ma	ma	linguistic		_	6:Guess=0-2	_	_
5	38-2	BO147	io	io	linguistic		_	6:Guess=0-2	_	_
6	38-3	BO147	da	da	linguistic		_	6:Guess=0-2	_	_
7	38-4	BO147	pallotti	pallotti	linguistic		_	6:Guess=0-8	_	_
8	38-5	BO147	ci	ci	linguistic		_	6:Guess=0-2	_	_
9	38-6	BO147	piglio	piglio	linguistic		_	6:Guess=0-6	_	_
10	38-7	BO147	le	le	linguistic		_	6:Guess=0-2	_	_
11	38-8	BO147	paste	paste	linguistic		_	6:Guess=0-5	_	_
12	38-9	BO147	>cioè:<	cioè	linguistic	Prolonged=3x1	_	2:Fast=0-4|6:Guess=0-4	_	_
13	38-10	BO147	ci	ci	linguistic		_	6:Guess=0-2	_	_
14	38-11	BO147	prend[o:	prendo	linguistic	ProsodicLink=Yes|Prolonged=5x1	_	6:Guess=0-6	5	5-6
15	38-12	BO147	mh	mh	linguistic		_	6:Guess=0-2	5	0-2
16	38-13	BO147	le	le	linguistic		_	6:Guess=0-2	5	0-2
17	38-14	BO147	brio~])	brio~	linguistic	Interrupted=Yes	End=52.997	6:Guess=0-5	5	0-5
18	39-0	BO146	[le	le	linguistic		Begin=51.927	_	5	0-2
19	39-1	BO146	lasagne]	lasagne	linguistic		End=52.887	_	5	0-7

INDICE

Applicare i principi FAIR al parlato

- ✓ **Findable:** Utilizzo di metadati ricchi e identificatori persistenti per rendere i dati parlati facilmente rintracciabili.
  - ➢ Tutti i moduli del KIParla e anche la risorsa nella sua interezza sono dotati di DOI
- ✓ **Accessible**: Definizione chiara delle condizioni di accesso, considerando le implicazioni etiche e legali.
  - ➢ Stiamo per depositare il KIParla su CLARIN-it
  - ➢ Possibile utilizzo di GitLab come repository di lavoro
- ✓ **Interoperable**: Adozione di formati standard e vocabolari condivisi per facilitare l'integrazione con altri dati e strumenti.
  - ➢ Alcuni standard presenti fin dall'inizio (Jefferson, NoSketchEngine), ma sono parzialmente indipendenti tra loro
  - ➢ ISO standard XML-TEI per il parlato > implica una riduzione di informazione a causa delle caratteristiche intrinseche del Jefferson
  - ➢ Sviluppo di un formato CoNLL – like
- ✓ **Reusable:** Documentazione dettagliata e licenze chiare per permettere il riutilizzo dei dati in diversi contesti.
  - ➢ La documentazione attuale non è *machine readable*, gli script non sono open source e non permettono la trasformazione autonoma dei dati

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.

- Il toolkit sviluppato può 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**

Rappresentare la soggettività, in modo trasparente

I nuovi formati multilivello permettono di mantenere **tracciabilità completa**: ogni modifica è registrata, ogni trascrittore e annotatore identificabile

- La gestione di versioni parallele consente di documentare il disaccordo e di confrontare approcci annotativi diversi.

visione "normativa" dell'annotazione  visione dialogica e pluralista

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.

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?