KIParla Corpus

Storia, scelte metodologiche, sfide future

Caterina Mauri*, Silvia Ballarè**

**Dipartimento di Lingue, Letterature e Culture Moderne*

***Dipartimento di Filologia Classica e Italianistica*

Indice

Indice

Background - i coordinatori del corpus KIParla

Caterina Mauri

- Tipologia linguistica;
- Mutamento linguistico;
- Variazione intra- e interlinguistica;
- Semantica e pragmatica.

Silvia Ballarè

- Sociolinguistica;
- Variazione intra- e interlinguistica;
- Contatto tra italiano e dialetti italo-romanzi.

Eugenio Goria

Massimo Cerruti

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Indice

Storia e contesto

Progetto SIR n. RBSI14IIG0 «LEADhoC - *The linguistic expression of ad hoc categories»*, 2015-2019 (PI Caterina Mauri)- www.leadhoc.org

2016: necessità di indagare la costruzione e la comunicazione di categorie ad hoc nel discorso parlato interazionale.

✓ Analisi dell'esistente: quali risorse, quali potenzialità

✓ Progetto di un nuovo corpus di italiano parlato

Corpora di italiano parlato: quadro generale

- ✓ Diversi corpora di parlato, costruiti per scopi di ricerca specifici, con metodologie diverse (es. map task, interviste, …) da singoli studiosi o da gruppi di ricerca
- ✓ … alcuni dei quali mai pubblicati

- ➢ **Parlaritaliano.it** raccoglie link a diverse risorse e strumenti per la trascrizione e annotazione della lingua parlata

Ulteriori mappature:

- https://accademiadellacrusca.it/it/contenuti/banche-dati-corpora-e-archivi-testuali/6228
- https://biblio.sns.it/it/corpora-della-lingua-italiana

- ✓ Nella progettazione del corpus KIParla abbiamo preso come riferimento i corpora di medie e grandi dimensioni, rappresentativi della variazione interna all'italiano parlato, che sono stati resi pubblicamente accessibili

Corpora di italiano parlato / VoLIP – Voce del LIP

Variazione diafasica, diatopica, diamesica

Dimensioni:

- ✓ 60h di conversazioni del corpus LIP, ca. 500.000 parole
- ✓ Dati raccolti all'inizio degli anni '90, trascritti, lemmatizzati, POS

Bilanciamento del campione:

- ✓ Quattro punti di raccolta: Milano, Firenze, Roma, Napoli
- ✓ Cinque categorie di situazioni comunicative

A) conversazioni faccia a faccia;
B) conversazioni telefoniche;
C) scambi comunicativi bidirezionali con alternanza di turno predefinita, come interviste, dibattiti, interazioni in aule scolastiche, esami orali, ecc.;
D) monologhi, come letture, sermoni, discorsi, ecc.;
E) programmi radiofonici e televisivi

Accessibilità:

- ✓ Liberamente accessibile su https://www.volip.it
- ✓ Trascrizioni e file audio allineati, no KWIC search

Metadati interrogabili:

Punto di raccolta (città), tipo di situazione comunicativa, sesso del parlante, genere testuale, grado di interazione, tipo di pianificazione, contesto sociale, struttura dell'evento comunicativo (mono/dialogica), canale

(Voghera et al. 2014)

Variazione diafasica, diamesica

Corpora di italiano parlato / LABLITA

Dimensioni:

- ✓ Ca. 200h di conversazioni, ca. 1.400.000 parole.
- ✓ Dati raccolti dagli anni ’70 fino agli anni ‘90

Bilanciamento del campione:

- ✓ Conversazione spontanea
- ✓ Varietà acquisizionale
- ✓ Trascrizione di film
- ✓ Radio e tv broadcast

Accessibilità: !

- ✓ Corpus non accessibile
- ✓ Parzialmente accessibile in una demo di C-ORAL-ROM: http://www.elda.org/en/proj/coralrom.html

Metadati interrogabili:

Canale, rapporto tra partecipanti, interazione libera o regolata, struttura dell’evento comunicativo (mono/dialogica), canale

(Cresti e Moneglia 2005)

Corpora di italiano parlato / CLIPS

Variazione diafasica, diamesica, diatopica

Dimensioni:

✓ 100h di conversazioni

Bilanciamento del campione:

✓ Quindici punti di raccolta, selezionati in base a criteri linguistici e socio-economici

✓ Quattro categorie di situazioni comunicative

Bari, Bergamo, Bologna, Cagliari, Catanzaro, Firenze, Genova, Lecce, Milano, Napoli, Palermo, Parma, Perugia, Roma, Venezia

A) parlato radiotelevisivo,
B) parlato dialogico *(map task, test delle differenze)*,
C) parlato letto,
D) parlato telefonico

Accessibilità: !

✓ Liberamente accessibile su http://www.clips.unina.it/it/corpus.jsp

✓ Trascrizioni e file audio, no KWIC search

Metadati interrogabili:

Punto di raccolta (città), tipo di situazione comunicativa, canale

(Sobrero e Tempesta 2007)

Variazione diafasica, diamesica, diatopica

Corpora di italiano parlato / Perugia Corpus

Dimensioni:

- ✓ Raccoglie corpora esistenti: LIP, sezione italiana del corpus Saccodeyl (Pérez-Paredes e Alcaraz Calero, 2009) e alcune parti del corpus CLIPS
- ✓ ca. 4.000.000 parole.

Bilanciamento del campione:

- ✓ Corpus di riferimento che include altri corpora esistenti
- ✓ Parlato spontaneo, parlato televisivo e radiofonico

Accessibilità: !

- ✓ Piattaforma in chiusura: https://www.unistrapg.it/cqpwebnew/index.php
- ✓ Nuova piattaforma (messaggio di errore): https://apps.unistrapg.it/cqpweb/

Metadati interrogabili:

??

(Spina 2014)

Per riassumere

! Problemi di **accessibilità e mantenimento**

- Interfacce di ricerca diverse: CQP o interfacce ad hoc;
- Corpora non facilmente accessibili (su supporti antiquati, o su server non attivi) – l'unico completamente accessibile è il VoLIP;
- KWIC search non sempre disponibile

= Problemi di **comparabilità e analisi**

- Dati poco integrabili, essendo stati raccolti sulla base di parametri ed esigenze scientifiche diversi;
- Audio non sempre accessibile;
- Non c'è connessione diretta tra trascrizione dato audio – tranne che nel VoLIP!

? Pochi **metadati relativi parlanti**

- Difficile considerare la variazione diatopica: punto di raccolta vs. origine dei parlanti;
- Quasi impossibile indagare la variazione diastratica;
- Rapporto tra interlocutori non noto – tranne che nel VoLIP!

Ideazione del corpus KIParla

KIParla nasce con lo scopo di integrare il panorama delle risorse esistenti con dati aggiornati, prendendo l'esperienza del VoLIP come modello. Il corpus KIParla intende offrire:

- ✓ un corpus **liberamente accessibile**
  che renda possibile e semplice la consultazione delle trascrizioni allineate con i file audio relativi

- ✓ un sistema di **metadatazione trasparente**
  sia rispetto alle caratteristiche oggettive delle situazioni comunicative che rispetto ai parlanti coinvolti

- ✓ un'**interfaccia di ricerca** basata su uno **standard internazionale**
  che offra funzioni di ricerca avanzate: KWIC, liste di frequenza, creazione di sottocorpora, ecc. (vd. §3)

- ✓ un'**infrastruttura replicabile, modulare e incrementale**
  che permetta l'espansione del corpus nel tempo attraverso una struttura modulare

Indice

Modularità e incrementalità

KIParla, un corpus modulare e incrementale:

- ✓ divisione interna in moduli indipendenti;
- ✓ possibilità di consultazione di ciascun modulo separatamente o di tutti i moduli in modo congiunto;
- ✓ possibilità di poter aggiungere nuovi moduli nel tempo.

➢ **Moduli**: corpora di italiano parlato distinti, accomunati da un insieme specifico di caratteristiche (nucleo minimo)

Risorsa multipla:

- ✓ struttura sottostante e infrastruttura di accesso comune ai diversi moduli;
- ✓ diversificazione interna relativa ai fattori extra-linguistici rilevanti e ai tipi di metadati disponibili per ciascun modulo, fermo restando il nucleo minimo.

→ Risposta all'esigenza di rendere comparabili dati che vengono raccolti in luoghi diversi, in fasi diverse, con scopi diversi e spesso con finanziamenti diversi.

Modularità e incrementalità

Nucleo minimo di caratteristiche perché un corpus di italiano parlato possa diventare un modulo del KIParla:

1. accessibilità ai metadati relativi ai parlanti
   - ➢ almeno rispetto a *età*, *provenienza* e *professione*
2. accessibilità ai metadati relativi alle situazioni comunicative
   - ➢ almeno rispetto a *luogo di raccolta* e *tipo di interazione*
3. utilizzo del software ELAN per la trascrizione (output .eaf, audio mp3)

✓ Nucleo minimo **necessario** per garantire la comparabilità tra moduli, ma **espandibile** per ottenere la quantità di informazioni e il livello di dettaglio richiesto dagli scopi di ricerca di ogni modulo.

Per ogni nuovo modulo del KIParla:

- **script** di conversione dal formato .eaf al formato di *NoSketch Engine*
- **spazio su server**,
- possibilità di **interrogazione congiunta** con gli altri moduli
- **supporto** nella fase di trascrizione e anonimizzazione dei dati, per omogeneità di trattamento.

Modularità e incrementalità

Corpora di piccole/medie dimensioni

- ✓ costruiti per scopi diversi, spesso complementari
- ✓ con dati di aree geografiche diverse e tipi di comunità diverse
- ➢ mantenendo però una **sostanziale comparabilità di fondo nella struttura e nell'accessibilità**.

- ➢ **La comparabilità e al tempo stesso la specificità di ogni modulo sono ciò che può rendere in futuro il corpus KIParla rappresentativo dell'italiano parlato**: quanti più moduli si aggiungeranno, tramite la collaborazione di diversi atenei, tante più dimensioni di variazione potranno essere esplorate.

Corpus design: il modulo KIP

Primo passo: **il modulo KIP**

- Parlato in ambiente universitario

→ Osservazione della dimensione di variazione diafasica (e diatopica)

✓ 2 punti di inchiesta: Bologna e Torino (2016-2019)
✓ Studenti e professori universitari
✓ Diversi contesti comunicativi

	Rapporto tra i partecipanti	Moderatore	Argomento
Conversazione libera	Simmetrico	Assente	Libero
Intervista semistrutturata	Simmetrico	Presente	Fisso
Ricevimento studenti	Asimmetrico	Assente	Libero
Esami	Asimmetrico	Assente	Libero
Lezioni	Asimmetrico	Presente	Fisso

Corpus design: il modulo ParlaTO

Secondo passo: **il modulo ParlaTO**

- Parlato della città metropolitana di Torino

→ Osservazione della dimensione di variazione diastratica (e diatopica)

✓ Un punto di inchiesta: Torino (2019)
✓ Parlanti con diversa caratterizzazione sociale
✓ Un unico contesto: l'intervista semistrutturata

	Fasce d'età
Giovani	18 ≤ x ≤ 30 anni
Adulti	30 < x ≤ 60 anni
Anziani	60 < x ≤ 89 anni

Corpus design: i due moduli

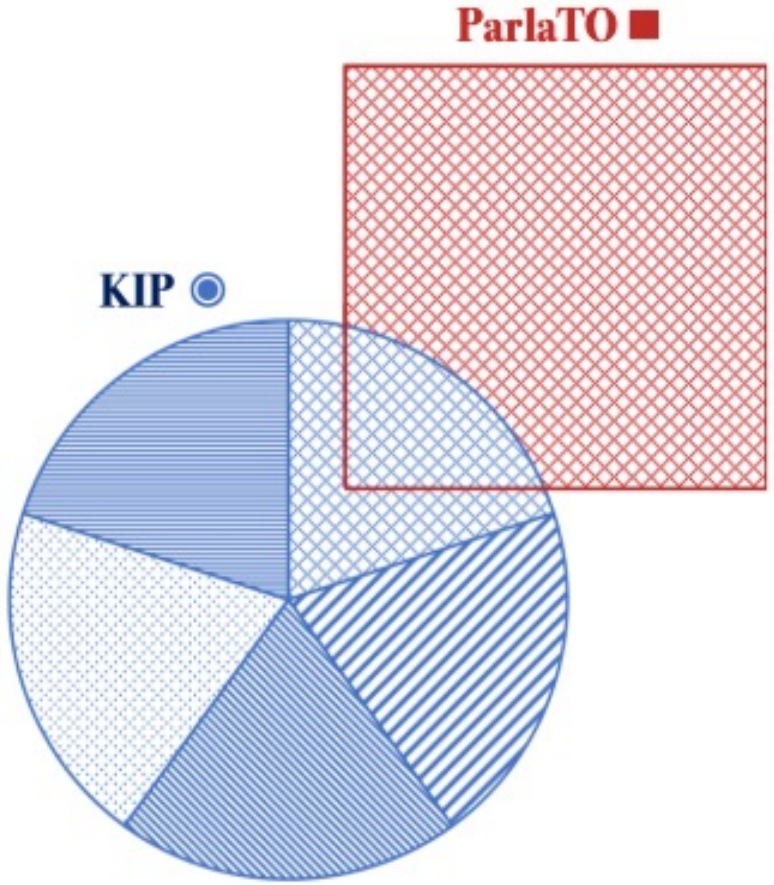
Conversazioni

Punto di raccolta:

- Bologna
- Torino

Tipo di interazione:

- Lezioni
- Esami
- Ricevimenti
- Interviste
- Conversazione spontanea



Parlanti

Età:

- $x \leq 35$
- $35 < x < 65$
- $x \geq 65$

Luogo di provenienza:

- Regioni italiane
- Paesi esteri

Titolo di studio:

- Licenza elementare/media
- Diploma tecnico/professionale
- Diploma di liceo
- Laurea
- Post-laurea

Raccolta e trascrizione: *it takes a village!*

2018 – oggi: più di 80 studenti e studentesse (delle università di Bologna e di Torino) hanno partecipato alla costruzione del corpus KIParla.

- Tirocini, tesi triennali, tesi magistrali, …

Formazione e supervisione

Aggiornamenti frequenti e incontri settimanali

Organizzazione e coordinazione

Raccolta dati: prima di iniziare

GDPR - Regolamento 2016/679

- Chi è responsabile dei dati?
- Quali metadati saranno conservati? Saranno aggregati?
- Dove saranno conservati i dati?
- Chi ha accesso hai dati?
- I contenuti delle registrazioni presentano informazioni sensibili (nomi e cognomi, indirizzi, …)?

Università degli Studi di Torino
Dipartimento di Studi Umanistici

Alma Mater Studiorum - Università di Bologna
Dipartimento di Lingue, Letterature e Culture Moderne

Informazioni sul trattamento dei dati personali ai sensi dell'art. 13 del Regolamento 2016/679/UE

Versione n. 1 del ___/___/2021

La parte istituzionale

Accordo di contitolarità per il trattamento dei dati personali

TRA

L'Alma Mater Studiorum – Università di Bologna, con sede in via Zamboni, n. 33, 40126 – Bologna, Italia, rappresentato legalmente dal Magnifico Rettore Giovanni Molari (di seguito definita "Università di Bologna" o "Contitolare del trattamento" o "Contitolare")

E

l'Università degli Studi di Torino, con sede legale in via Verdi, n. 8, 10124 - Torino, Italia, rappresentata legalmente dal Magnifico Rettore Stefano Geuna (di seguito definita, "contraente" o "Contitolare del trattamento" o "Contitolare");

ACCORDO DI COLLABORAZIONE

TRA

Il **Dipartimento di Studi Umanistici dell'Università di Torino**, con sede legale a Torino, in via Sant'Ottavio 20 – C.F. 80088230018, P.I. 02099550010, rappresentato da

- il Direttore Prof. Donato Pirovano, nato a Como il 17/2/1964, autorizzato alla stipula del presente atto con delibera del Consiglio di Dipartimento del 15/11/2021.
- la Dott.ssa Antonella Trombetta - Direttrice della Direzione Ricerca e Terza missione, nata a Torino il 6/10/1970, per quanto di competenza e per quanto previsto dagli artt. 29 comma 1 e 66 comma 1 del Regolamento di Amministrazione, Finanza e Contabilità emanato con Decreto Rettorale n. 3106 del 26/09/2017 che dispone in ordine alla capacità negoziale e alla stipulazione del contratto,

E

Il **Dipartimento di Lingue, letterature e culture moderne** dell'Università di Bologna con sede legale a Bologna, in via Cartoleria 5 – C.F. 80007010376, P.I. 01131710376, rappresentato da

- il Direttore Maurizio Ascari, nato a Loiano (BO) il 06/06/1965, autorizzato alla stipula del presente atto con delibera del Consiglio di Dipartimento del 11/11/2021 domiciliato, ai fini del presente atto, presso la sede del Dipartimento.

Il **Dipartimento di Filologia classica e italianistica dell'Università di Bologna,** con sede legale a Bologna, in via Zamboni 32 – C.F. 80007010376, P.I. 01131710376, rappresentato da

- il Direttore Prof. Nicola Grandi, nato a Ferrara il 30/08/1973, autorizzato alla stipula del presente atto con delibera della Giunta di Dipartimento del 15/11/2021 domiciliata, ai fini del presente atto, presso la sede del Dipartimento

Raccolta dati: strumenti

Moduli:

- Liberatoria;
- Raccolta dei metadati.

Registrazione:

- Registratore Zoom H4n Pro;
- Smartphone (per ragioni pratiche).

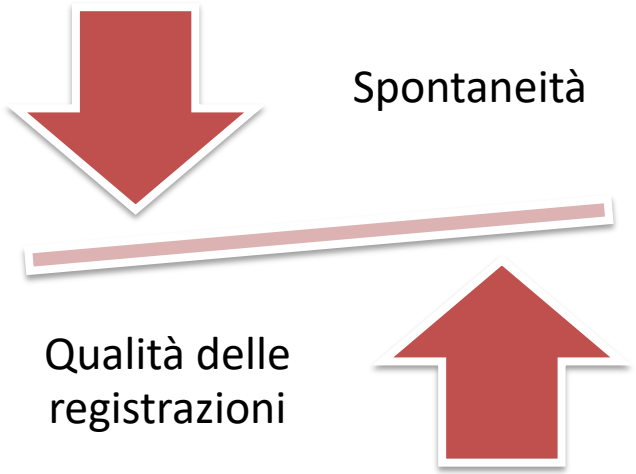
Canovaccio per l'intervista (Labov, 1984: 32-33)

- to elicit narratives of personal experience, where community norms and styles of personal interaction are most plainly revealed, and where style is regularly shifted towards the vernacular;
- to stimulate group interaction among the people present, and so record conversation not addressed to the interviewer;
- to isolate from a range of topics those of greatest interest to the speaker, and allow him or her to lead in defining the topic of conversation.

Raccolta dati: impostazione

La maggior parte dei dati sono stati raccolti dagli studenti e dalle studentesse delle università di Bologna e di Torino.

- Formazione (metodologia della raccolta dati);
- Selezione dei partecipanti.



Raccolta dati: archiviazione

Dopo ogni registrazione, utilizzando la cartella *OneDrive*, il raccoglitore ha:

- Caricato e denominato (con un codice alfanumerico) il file audio (.mp3, .wav);
- Inserito (su due files excel distinti) informazioni riguardo ai partecipanti e alla conversazione;
- Caricato le liberatorie firmate dai partecipanti;
- Avviato la trascrizione.

Conversazioni – modulo KIP

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	type	code	duration	icipants nur	p1	p2	p3	p4	p5	p6	p7	pants relati	moderator	topic	year	point
33	conversazione libera	BOA3017	0:30:22	4	BO139	BO145	BO146	BO147				simmetrico	no	libero	2019	BO
34	conversazione libera	BOA3018	0:27:14	2	BO139	BO145						simmetrico	no	libero	2019	BO
35	conversazione libera	BOA3019	0:23:15	4	BO149	BO150	BO151	BO152				simmetrico	no	libero	2019	BO
36	conversazione libera	BOA3020	0:22:52	1	BO152	BO153	BO154					simmetrico	no	libero	2019	BO
37	conversazione libera	BOA3021	1:11:38	4	BO155	BO156	BO157	BO158				simmetrico	no	libero	2019	BO

Partecipanti – modulo ParlaTO

	A	B	C	D	E	F	G	H
1	participant code	participant occupa	participant sex	files in which participant appears	participant bi	participant ag	participant degree	
2	TOR001	intell	F	PTA002	lombardia	26-30	phd	
3	TOR002	intell	M	PTA002	piemonte	26-30	phd	
4	TOR003	intell	M	PTA002	piemonte	26-30	laurea	
5	TOI002	oper	M	PTA002	piemonte	26-30	it	
6	TOI003	oper	M	PTA002	piemonte	26-30	it	

Trascrizione dei dati

La maggior parte dei dati sono stati trascritti dagli studenti e dalle studentesse delle università di Bologna.

- Formazione;
- Monitoraggio periodico.

È stato utilizzato il softwer ELAN. 

Per dare conto di aspetti conversazionali (sovrapposizioni, comportamenti non verbali, …) si è impiegata una semplificazione del sistema proposto da Jefferson (2004).

"Nearly-globalized set of instructions for transcription" (Slembrouck 2007: 823).

,	Intonazione ascendente	<ciao>	Pronuncia (più) lenta
.	Intonazione discendente	[ciao]	Sovrapposizioni tra parlanti
:	Suono prolungato	(ciao)	Testo di difficile comprensione
(.)	Pausa breve	xxx	Testo non comprensibile
>ciao<	Pronuncia (più) veloce	((ride))	Comportamento non verbale

Uniformazione e anonimizzazione dei dati

Tutte le trascrizioni sono poi state controllate (in termini di uniformazione e anonimizzazione) da un'unica persona.

Revisione:

- Refusi;
- Sovrapposizioni.
- …

Anonimizzazione:

- Cancellazione dati sensibili dalla trascrizione e dalla traccia audio.

Il risultato / modulo KIP

Attività	**Bologna**	**Torino**	*TOT*
Conversazioni libere	10:00:37	06:22:24	16:23:01
Esami	03:09:34	03:10:48	6:20:22
Lezioni	12:19:39	13:25:33	25:45:12
Interviste semistrutturate	06:18:37	07:47:38	14:06:15
Ricevimento studenti	02:59:11	03:49:08	6:48:19
TOT	34:47:38	34:35:30	**69:23:08**
Informatori	150	123	273
Token	329.464	331.711	661.175

Il risultato / modulo KIP

Parlanti	**Classe d'età:**	under25, 26-30, 31-35, 36-40, 41-45, 46-50, 51-55, 56-60, over60
	Sesso:	M, F
	Regione delle scuole superiori:	Abruzzo, Basilicata, Calabria, Campania, Emilia-Romagna, Estero, Friuli-Venezia Giulia, Lazio, Liguria, Lombardia, Marche, Molise, Piemonte, Puglia, Sardegna, Sicilia, Toscana, Trentino-Alto Adige, Umbria, Valle d'Aosta, Veneto
	Occupazione:	p (professore), s (studente)
Interazioni	**Tipo:**	conversazione libera, esami, interviste semistrutturate, lezioni, ricevimento studenti
	Luogo:	Bologna, Torino
	Relazione:	Simmetrica, asimmetrica
	Partecipanti:	1, 2, 3, 4, 5, 6
	Moderatore:	Sì, no
	Topic:	Fisso, libero

Il risultato / modulo ParlaTO

	Giovani (18 ≤ x ≤ 30 anni)	**Adulti** (30 < x ≤ 60 anni):	**Anziani** (60 < x ≤ 89 anni):	*TOT*
	17:33:20	14:49:53	16:15:31	**48:38:44**
Informatori	35	28	25	88
Token	258.083	145.425	131.173	552.461

Il risultato / modulo ParlaTO

Parlanti	**Classe d'età:**	16-20, 21-25, 26-30, 31-35, 36-40, 41-45, 46-50, 51-55, 56-60, 61-65, 66-70, 71-75, 76-80, 81-85, over85
	Sesso:	M, F
	Regione di nascita:	Abruzzo, Basilicata, Friuli-Venezia Giulia, Lazio, Lombardia, Piemonte, Puglia, Sardegna, Sicilia, Trentino-Alto Adige, Veneto
	Titolo di studio:	elem (diploma elementare), medie (licenza media), it (diploma di istituto tecnico o professionale), lic (diploma di liceo), laurea (laurea triennale, magistrale e a ciclo unico), phd (dottorato di ricerca)
	Occupazione:	artig (artigiani, operai specializzati e agricoltori), comm (professioni qualificate nelle attività commerciali e nei servizi), disocc (disoccupati), impr (legislatori, imprenditori e alta dirigenza), intell (professioni intellettuali, scientifiche e di elevata specializzazione), nonq (professioni non qualificate), oper (conduttori di impianti, operai di macchinari fissi e mobili e conducenti di veicoli), pens (pensionati), stud (studenti)
Interazioni	**Partecipanti:**	2, 3, 4, 5, 6
	Lingue:	italiano, italiano e dialetto

Il risultato / KIParla

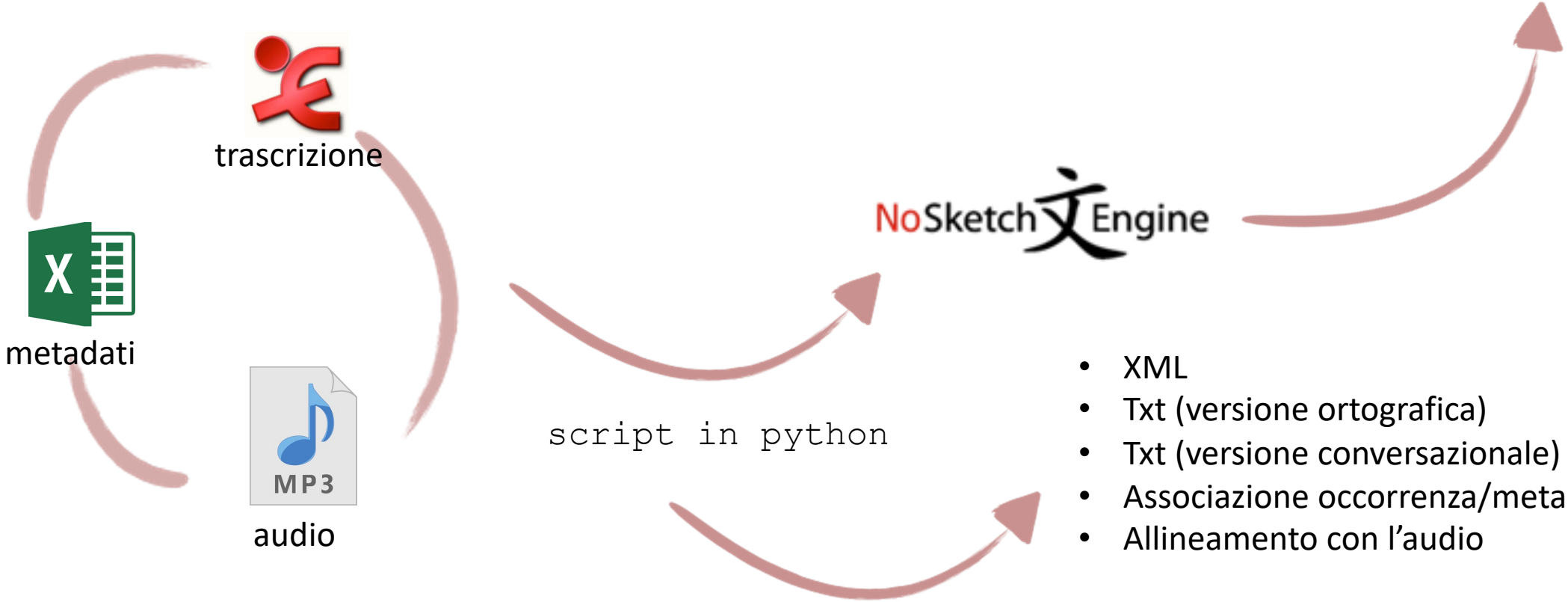
Ore	110:14:14
Informatori	340
Tokens	1.125.996

Il corpus è interrogabile anche congiuntamente utilizzando come filtri di ricerca quelli condivisi dai due moduli, ovvero:

- tipo di conversazione;
- argomento della conversazione;
- presenza di un moderatore;
- numero di partecipanti;
- anno in cui è stata registrata la conversazione;
- punto in cui è stata registrata la conversazione;
- occupazione del partecipante;
- età del partecipante;
- provenienza del partecipante.

Pubblicazione dei dati

- Standard internazionale
- Opportunità di funzionalità avanzate
- Open source



- XML
- Txt (versione ortografica)
- Txt (versione conversazionale)
- Associazione occorrenza/metadati
- Allineamento con l'audio

Accessibilità dei dati

- Trascrizioni (ortografiche e conversazionali):

  consultabili **liberamente** attraverso la piattaforma NoSketch Engine;

  è possibile scrivere ai coordinatori del corpus per ricevere i files .txt delle conversazioni.

- Audio:

  previa registrazione, è possibile ascoltare le tracce audio on line (a partire da NoSketch Engine);

  non è possibile ricevere le tracce audio (G.D.P.R.).

www.kiparla.it

Indice

Usare il corpus / esempio 1

Tipi di ricerche:

- ✓ Query types
- ✓ Context
- ✓ Text types ⟹ cambiano i filtri di ricerca a seconda del modulo selezionato

Altre funzionalità:

- ✓ Creazione di sottocorpus
- ✓ Liste di frequenza (Word list, n-grams, liste di frequenza per ogni metadato)
- ✓ Corpus info

www.kiparla.it

Usare il corpus / esempio 2

KWIC:

- ✓ Save concordance (txt, csv, xml)
- ✓ Liste di frequenza (Word list, n-grams, liste di frequenza per ogni metadato)
- ✓ View options: si possono selezionare diversi metadati da visualizzare ed esportare insieme alle occorrenze
- ✓ Sort
- ✓ Random sample
- ✓ Filtri ulteriori
- ✓ Analisi di frequenza delle occorrenze (in relazione ai diversi metadati)
- ✓ Collocation candidates

www.kiparla.it

Usare il corpus / esempio 3

Metadati e link:

- ✓ Click su KWIC > espansione del contesto
- ✓ Click sul codice conversazione
  - ➢ Tutti i metadati relativi al parlante e alla conversazione
  - ➢ Link alla conversazione in formato html (ortografica)
  - ➢ Link alla conversazione in formato html (conversazionale)
  - ➢ Link al file audio allineato alla specifica occorrenza (3 sec. prima)

www.kiparla.it

Indice

Prossimi passi e sfide future

Sviluppi computazionali:

- ✓ Lemmatizzazione
- ✓ POS tagging
- ✓ Ulteriori livelli di annotazione
- ✓ Nuova versione di NoSketch Engine

BOSCO, Cristina et al. 2020.
KIPoS @ EVALITA2020: Overview of the Task on KIParla Part of Speech Tagging In: *EVALITA Evaluation of NLP and Speech Tools for Italian - December 17th, 2020: Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian Final Workshop* [online]. Torino: Accademia University Press, 2020. Available on the Internet: <http:// books.openedition.org/aaccademia/7743>. ISBN: 9791280136329. DOI: https://doi.org/10.4000/ books.aaccademia.7743.

Nuovi moduli e nuove collaborazioni:

- ✓ ParlaBZ
- ✓ KIPasti
- ✓ ParlaBO
- ✓ …

➢ Il corpus KIParla continua a crescere, all'insegna della **sostenibilità** (anche in assenza di finanziamenti esterni)

Nuovi moduli / ParlaBZ

Responsabili: Daniela Veronesi, Alex Piovan (Libera Università di Bolzano)

Caratteristiche: parlato spontaneo raccolto in contesti diversi (interviste semistrutturate, cene) con parlanti di diversa caratterizzazione sociale nella città di Bolzano.

Le trascrizioni sono state effettuate con particolare attenzione ad aspetti di natura conversazionale

- ✓ Un unico punto di inchiesta: Bolzano
- ✓ Parlanti con diversa caratterizzazione sociale
- ✓ Diversi contesti comunicativi

Dati:

- ➢ ore: 5:15:56;
- ➢ numero di registrazioni: 10;
- ➢ numero di parlanti: 16.

→ Osservazione della variazione diastratica e diafasica

Stato attuale: in fase di pubblicazione.

Nuovi moduli / KIPasti

Responsabili: Caterina Mauri, Silvia Ballarè (Università di Bologna)

Caratteristiche: parlato spontaneo raccolto durante pasti (pranzi, cene, …) nelle diverse aree geografiche d'Italia (campione rappresentativo delle macroaree di nord, centro, sud e isole).

- ✓ Diversi punti di inchiesta: nord, centro, sud
- ✓ Parlanti con diversa caratterizzazione sociale
- ✓ Un unico contesto: la conversazione spontanea a tavola

Dati (provvisori):

- ➢ ore: 41:30:58;
- ➢ numero di registrazioni: 63;
- ➢ numero di parlanti: 163.

→ Osservazione della variazione diastratica e diatopica

Stato attuale: in fase di trascrizione.

Nuovi moduli / ParlaBO

Responsabili: Caterina Mauri, Silvia Ballarè (Università di Bologna)

Caratteristiche: interviste semistrutturate registrate nella città metropolitana di Bologna a parlanti con diversa caratterizzazione sociale. Il campione, analogamente a quanto fatto per il ParlaTO, è bilanciato per fasce di età.

- ✓ Un unico punto di inchiesta: Bologna
- ✓ Parlanti con diversa caratterizzazione sociale
- ✓ Un unico contesto: intervista semi-strutturata

Dati (provvisori):

- ➢ ore: 42:56:37;
- ➢ numero di registrazioni: 57;
- ➢ numero di parlanti: 99.

→ Osservazione della variazione diastratica (e diatopica)

Stato attuale: in fase di raccolta e trascrizione.

ATTENZIONE!

- ✓ **Normativa sulla privacy**: cambiamenti della normativa in itinere!

  → Le persone registrate devono essere contattabili nel medio periodo, per aggiornamenti relativi alle liberatorie (es. contatto mail)

- ✓ **Documentazione istituzionale** (ufficio legale e ufficio privacy)

  → Tempi molto lunghi, procedure diverse da ateneo ad ateneo, numerosi scambi necessari per trovare un accordo relativo al trattamento e alla titolarità dei dati (soprattutto nei progetti che coinvolgono più atenei)

- ✓ **Sostenibilità costi del server**

  → Molte risorse sono destinate a scomparire gradualmente dopo la fine dei progetto all'interno del quale sono nate: una volta finito il finanziamento posso ancora pagare per il server? Cercare soluzioni sostenibili nel tempo, presidiate da personale competente, che aggiorni il server e lo renda *compliant* con le normative – che anche in questo caso cambiano molto velocemente

ATTENZIONE!

✓ **Sostenibilità del processo**

→ Forme di collaborazione sostenibili: es. tirocini (CFU in cambio di collaborazione), laboratori, …

✓ **Tipo di corpus**: corpus di interesse generale o corpus per scopi specifici?

→ Scelte metodologiche diverse a seconda degli scopi.

- Per la ricerca fonetico-prosodica, il file audio deve essere di alta qualità e le registrazioni vanno fatte necessariamente con microfoni appositi (no telefoni! Problema: meno naturalezza).
- Per la ricerca conversazionale, la trascrizione deve essere seguire le regole di trascrizione complete e deve idealmente essere accompagnata da materiale video (problema: più tempo di elaborazione dei dati, riduzione del campione finale, problemi di privacy ulteriori per i dati video).
- Per la ricerca sociolinguistica, vanno curati i metadati relativi agli assi di variazioni principali (problema: per un campione rappresentativo servono grandi moli di dati).
- ….

ATTENZIONE!

- ✓ **Accessibilità**: chi può accedere a quali dati?
  - → È necessario prevedere la modalità di registrazione/monitoraggio del traffico sul corpus, per produrre un elenco aggiornato di persone che hanno accesso ai dati audio (e video). Soluzioni possibili: registrazione tramite username/password individuali, registrazione tramite strumenti come la newsletter
  - → I dati relativi a minori (audio/video) presentano maggiori problemi relativi alla privacy, quindi occorre prevedere molto tempo per predisporre la documentazione istituzionale e legale.

- ✓ **Interfaccia di ricerca**
  - → L'uso di standard internazionali (come NoSketch Engine) ha molti vantaggi, ma ci sono dei requisiti di input, quindi va prevista fin dall'inizio la predisposizione di script appositi

- ✓ **Collaborazione informatica** che garantisca continuità nel tempo
  - → Aggiornamento script, soluzione a eventuali problemi di natura informatica

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Grazie!

caterina.mauri@unibo.it
silvia.ballare@unibo.it

www.unibo.it