DIPARTIMENTO DI LINGUE, LETTERATURE E CULTURE MODERNE
ALMA MATER STUDIORUM • UNIVERSITÀ DI BOLOGNA

VIA CARTOLERIA, 5 – 40124 BOLOGNA – ITALIA

Attività di tirocinio nell'ambito del progetto
'Corpus di italiano parlato KIParla'

Bologna, 18 febbraio 2024

1. Presentazione del progetto 'Corpus di italiano parlato KIParla'

Il corpus di italiano parlato KIParla (www.kiparla.it) è stato concepito all'interno del progetto SIR (n. RBSI14IIG0, www.leadhoc.org) 'LEAdhoC - *Linguistic expression of ad hoc,* presso l'Università di Bologna al Dipartimento di Lingue, Letterature e Culture Moderne, sotto il coordinamento della Prof.ssa Caterina Mauri. Il progetto LEAdhoc si è concluso nel settembre 2019, ma la costruzione del corpus KIParla è continuata, grazie all'aggregazione di docenti e ricercatori da diverse università italiane, in primo luogo le Università di Bologna e Torino. Dal 2019 i coordinatori del Corpus KIParla sono Caterina Mauri e Silvia Ballarè (Università di Bologna) con Eugenio Goria e Massimo Cerruti (Università di Torino). Al momento, l'ampliamento del corpus KIParla è l'obiettivo principale del progetto PRIN 2022 PNRR 'DiverSIta – *Diversity in Spoken Italian*' (n. P2022RFR8T, 2023-2025, https://site.unibo.it/divers-ita/it), coordinato dalla Prof.ssa Caterina Mauri, che vede coinvolti i due atenei di Bologna e Torino.

Il corpus KIParla è una risorsa innovativa, disponibile gratuitamente per tutti coloro che lavorano sull'italiano parlato, è ricercabile tramite l'interfaccia No Sketch Engine, ed è strutturata in modo da permettere uno sviluppo incrementale e modulare nel tempo.

Attualmente, il corpus KIParla consiste di due moduli già pubblicati (KIP e ParlaTO), tre moduli in via di pubblicazione (ParlaBZ, ParlaBO e KIPasti), e quattro moduli in via di realizzazione (ParlaNA, ParlaMI, Stra-ParlaBO, Stra-ParlaTO, vd. sotto). Nel tempo, si prevede che il corpus KIParla possa crescere ulteriormente tramite collaborazioni con progetti esistenti e progetti futuri, che andranno a costituire nuovi moduli del corpus.

Moduli pubblicati

- Il primo modulo (KIP, http://kiparla.it/kiparla-bo-to/) è ormai concluso, grazie anche al supporto fornito dagli studenti che hanno partecipato al precedente tirocinio LEADhoC (con cui il presente progetto è in piena continuità), e conta circa 70 h di conversazioni registrate a Bologna e Torino. La sua principale caratteristica è quella di contenere campioni di vari tipi di interazione osservabili in ambito universitario, in particolare lezioni, ricevimento studenti, esami, conversazione libera fra studenti, interviste semistrutturate a studenti. È dunque garantita una certa omogeneità degli intervistati per quanto riguarda il profilo socio-culturale (studenti e docenti), mentre si è cercato di avere la massima differenziazione possibile per quanto riguarda i tipi di interazione osservati, che si distinguono per i diversi livelli di formalità, per il carattere monologico o dialogico e per il tipo di rapporto che intercorre tra i partecipanti. Il corpus è online, ed è liberamente accessibile e ricercabile.

- Anche il secondo modulo (ParlaTO, http://kiparla.it/parlato/) è concluso, grazie alla collaborazione di

ricercatori bolognesi e torinesi e al finanziamento ottenuto dalla Cassa di Risparmio di Torino. L'obiettivo del ParlaTO è quello di restituire un'immagine della realtà plurilingue torinese contemporanea in tutte le sue componenti. Le produzioni linguistiche presenti nel corpus sono raccolte per mezzo di interviste semi-strutturate individuali e di gruppo e hanno come argomento principale la città di Torino e delle aree circostanti. Le interazioni registrate, oltre che essere materiale utile a indagini di natura linguistica, offrono un'immagine contemporanea e dinamica del territorio che si racconta attraverso le voci degli intervistati.

Moduli conclusi o in fase di revisione

- Il modulo ParlaBZ contiene interviste e cene registrate e trascritte a Bolzano, grazie alla collaborazione con la Prof.ssa Daniela Veronesi dell'Università di Bolzano. I dati sono già stati revisionati e anonimizzati, l'accordo inter-ateneo per il trattamento dei dati è stato firmato, quindi il modulo sarà pubblicato nei primi mesi del 2024.

- Il modulo KIPasti consiste in circa 40 ore di parlato spontaneo registrato durante pranzi e cene in famiglia o con amici, cioè in situazioni in cui i parlanti condividono un contesto e un background comune. Le conversazioni includono fino a 4 partecipanti (per lo più 2 o 3) estremamente differenziati per età, profilo educativo, professione e repertorio. Il modulo è stato bilanciato in base all'area geografica di raccolta dei dati (Nord, Centro, Sud e isole) e mostra quindi una grande variazione diatopica. La situazione comunicativa specifica è in questo caso caratterizzata da un registro informale, dall'assenza di un argomento predefinito e dalla presa di turno libera. Il modulo è stato coordinato da Caterina Mauri e Silvia Ballarè ed è al momento in via di revisione. Si prevede la pubblicazione nei primi mesi del 2024.

- Il modulo ParlaBO è composto da circa 50 ore di interviste semi-strutturate raccolte nell'ambiente urbano della città di Bologna. Rispecchia la struttura del modulo parallelo raccolto nella città di Torino (ParlaTO). Il dataset è stato bilanciato in base all'età dei partecipanti, in modo da essere rappresentativo di diverse generazioni, e comprende parlanti con profili socioculturali e repertori diversi, rendendo conto della realtà plurilingue che ne caratterizza il tessuto sociale, dando voce a tutte le sue componenti. Le interviste semi-strutturate hanno come argomento principale la città di Bologna e delle aree circostanti, le biografie linguistiche dei cittadini bolognesi e il loro rapporto con il quartiere, il loro rapporto con le tradizioni locali. Il modulo è stato coordinato da Caterina Mauri e Silvia Ballarè ed è al momento in via di conclusione, è terminata la raccolta dati e stiamo procedendo con la trascrizione e revisione dei dati. Si prevede la pubblicazione entro la metà del 2024.

Moduli in costruzione

- Il modulo Stra-ParlaBO sarà composto da circa 50 ore di dati orali di interviste semi-strutturate e parlato spontaneo (ad esempio conversazioni a tavola) tra parlanti con trascorso di migrazione internazionale (Speakers with International Migration Background, SIMB) che vivono nell'area urbana di Bologna. Saranno coinvolte persone provenienti da quattro comunità linguistiche (cinese, bengalese, arabo marocchina, ucraina) e i partecipanti varieranno in base al paese di origine, alla L1, all'età, al livello di istruzione, al tempo trascorso in Italia e al tipo di occupazione. Il modulo verrà realizzato dall'Unità di Ricerca di Bologna all'interno del progetto PRIN 2022 PNRR (DiverSIta, https://site.unibo.it/divers-ita/it/partecipanti/unita-di-ricerca-di-bologna).

- Il modulo STRA-ParlaTO sarà composto da circa 50 ore di dati orali di interviste semi-strutturate e parlato spontaneo (ad esempio conversazioni a tavola) tra parlanti con trascorso di migrazione internazionale (Speakers with International Migration Background, SIMB) che vivono nell'area urbana di Torino. Saranno coinvolte persone provenienti da quattro comunità linguistiche e i partecipanti varieranno in base al paese di origine, alla L1, all'età, al livello di istruzione, al tempo trascorso in Italia e al tipo di occupazione. Il modulo verrà realizzato dall'Unità di Ricerca di Torino all'interno del progetto PRIN 2022 PNRR (DiverSIta, https://site.unibo.it/divers-ita/it/partecipanti/unita-di-ricerca-di-torino).

- Il modulo ParlaNA comprende dati raccolti nel territorio di Napoli, in diverse situazioni comunicative, ed è coordinato dalla Prof.ssa Margherita Di Salvo (Università di Napoli Federico II). Il modulo è in costruzione e l'accordo inter-ateneo per il trattamento dei dati è stato firmato.

- Il modulo ParlaMI comprende interviste semi-strutturate e conversazioni a tavola registrate nel territorio di Milano. Il corpus verrà stato bilanciato in base all'età dei partecipanti e al tipo di interazione, in modo da essere rappresentativo di diverse generazioni e diverse situazioni comunicative, e comprende parlanti con profili socioculturali e repertori diversi, rendendo conto della realtà plurilingue che caratterizza il tessuto sociale di Milano, dando voce a tutte le sue componenti. Le interviste semi-strutturate hanno come argomento principale la città di Milano e delle aree circostanti, le biografie linguistiche dei cittadini milanesi, il loro rapporto con il quartiere e con le tradizioni locali. Il modulo è coordinato da Caterina Mauri e Silvia Ballarè (Università di Bologna) e Federica Da Milano (Università di Milano Bicocca). È stata attivata la convenzione per l'apertura del tirocinio agli studenti milanesi e il modulo è nelle prime fasi di costruzione.

Aspetti innovativi del corpus KIParla

I principali aspetti innovativi di rispetto alle risorse attualmente esistenti sono:

- Un sistema di metadatazione che permette, se pure in forma anonima, di risalire alle caratteristiche sociolinguistiche del parlante (età, provenienza, titolo di studio)
- L'adozione di una liberatoria che autorizza all'utilizzo e alla diffusione dei dati in forma anonima
- L'allineamento sistematico della trascrizione con l'audio delle registrazioni, in modo che gli utenti possano risalire direttamente al dato reale e non semplicemente alla trascrizione
- La natura incrementale e modulare della risorsa.

Accanto alla costruzione del corpus e all'ampliamento dei suoi moduli, il progetto prevede anche un monitoraggio costante dei dati che emergono dal tessuto plurilingue di due città socialmente complesse come Bologna e Torino. Nell'ambito dell'analisi dei dati, il progetto prevede **analisi di corpora plurilingui** e **analisi di parlato in altri contesti europei ed extra-europei**.

---

2. Progetto di Tirocinio

La natura spiccatamente applicativa delle attività necessarie alla realizzazione del *corpus* richiede la messa in

atto di competenze altamente interdisciplinari, applicabili - oltre che nella ricerca di base in scienze umane - anche in contesti lavorativi esterni all'Università e in particolare nel settore dei servizi, della gestione di database linguistici (e non) e dell'informatica umanistica. La partecipazione al progetto costituisce dunque un'importante occasione formativa per gli studenti del Dipartimento di Lingue, Letterature e Culture Moderne, nonché degli altri Dipartimenti che aderisocno al tirocinio.

Per questa ragione, anche considerando la natura modulare, incrementale e autosostenibile del corpus KIParla, il progetto è aperto al tirocinio a tempo indeterminato agli studenti dei corsi di Laurea in *Lingue e Letterature Straniere* (250 ore per 9 cfu), *Lingue, Mercati e Culture dell'Asia* (250 ore per 9 cfu), *Language, society and communication* (180 ore per 6 cfu), *Lingua e cultura italiane per stranieri* (180 ore per 6 cfu), *Dati, metodi e modelli per le scienze linguistiche* (180 ore per 6 CFU) per tre diverse tipologie di attività:

- **Raccolta dati sul campo (italiano e lingue e lingue straniere presenti nella città di Bologna)**
  - ✓ MANSIONI PRINCIPALI: i tirocinanti, singolarmente o in gruppi, raccoglieranno dati di parlato, individuando situazioni comunicative adatte e soggetti da coinvolgere nelle registrazioni; saranno intervistati anche parlanti di origine straniera, in modo tale da raccogliere dati di parlato di italiano come L2 affiancato da altre lingue europee ed extraeuropee (ad es. spagnolo, varietà di arabo e cinese). Inoltre, si occuperanno di fare compilare il consenso informato e gestiranno i metadati dei partecipanti.
  - ✓ OBIETTIVI FORMATIVI: acquisizione di competenza nella realizzazione delle principali tipologie di intervista in uso nelle scienze sociali; gestione di archivi audio e video utilizzando software innovativi nel campo dell'informatica umanistica e dell'elaborazione del linguaggio naturale (NLP); gestione di database; acquisizione di conoscenze di base relative alle norme vigenti in materia di *privacy* (GDPR).

- **Trascrizione del parlato**
  - ✓ MANSIONI PRINCIPALI: i tirocinanti parteciperanno alla trascrizione delle interviste secondo il sistema Jefferson, applicato all'italiano e alle lingue straniere
  - ✓ OBIETTIVI FORMATIVI: acquisizione di competenza nelle principali convenzioni di trascrizione del parlato; utilizzo del software ELAN, impiegato anche nella produzione di sottotitoli; trascrizione di varietà parlate di lingue europee e extra-europee.

- **Preparazione dei dati per il trattamento automatico**
  - ✓ MANSIONI PRINCIPALI: i tirocinanti predisporranno le interviste trascritte per l'inserimento sulla piattaforma NoSketchEngine, si occuperanno della codifica dei dati in formato XML, parteciperanno alla creazione di una treebank gold sui dati del KIParla, idealmente da far confluire nelle treebank UD per l'italiano. Se necessario si relazioneranno con il personale tecnico per il caricamento dei dati.
  - ✓ OBIETTIVI FORMATIVI: conoscenza dei principali *corpora* di parlato attualmente consultabili; familiarizzazione con il sistema XML e il sistema di codifica di dati del parlato; conoscenza dei principali tipi di *query*.

- **Focus sulle lingue straniere presenti nella città di Bologna, Torino e Milano: analisi approfondite di singole lingue**
  - ✓ MANSIONI PRINCIPALI: i tirocinanti effettueranno analisi approfondite di singole lingue, che varieranno in relazione alle loro competenze e alle lingue di migrazione presenti nei dati;

ricorreranno a diverse risorse linguistiche e corpora di parlato per le lingue in questione, sia per monitorare l'uso di specifiche costruzioni, che per individuare fenomeni di contatto nel sistema e nel discorso.

- ✓ Si prevede un interesse particolare per le lingue più rappresentate nei contesti di migrazione di Torino e Bologna, anche come lingue veicolari (es. spagnolo, francese, cinese, arabo, russo, ucraino, bengalese), senza che questo precluda, tuttavia, l'inclusione di lingue ulteriori.
- ✓ Il compito prevede la collaborazione sistematica fra tirocinanti con competenze diverse. Per questo motivo si invita la partecipazione di studenti specializzati sia in lingue europee, sia in lingue dell'Asia e dell'Africa.
- ✓ OBIETTIVI FORMATIVI: i tirocinanti acquisiranno competenza nei principali strumenti di analisi specifici per singole lingue, quali *corpora* e archivi di parlato, impareranno a formulare autonomamente *queries* di diversa complessità.

Durante il tirocinio gli studenti saranno formati da un tutor (membro del progetto KIParla) per lo svolgimento delle seguenti attività:

- Realizzazione dei principali tipi di intervista in uso nelle scienze sociali
- Realizzazione di registrazioni con registratore professionale e gestione dei file audio
- Utilizzo dei principali *software* per la trascrizione e l'analisi del parlato
- Creazione di un sistema di annotazione
- Costituzione e gestione dei *corpora*
- Ricerche specifiche all'interno di corpora di lingue parlate europee e extra-europee

La formazione dei tirocinanti si terrà presso il Dipartimento LILEC, via Cartoleria 5, Bologna e attraverso la piattaforma Teams, che verrà usata settimanalmente per le riunioni del tirocinio.

A ogni tirocinante verranno assegnati dei compiti relativi a uno o più degli ambiti sopra descritti, sulla base di eventuali competenze pregresse, degli interessi specifici del candidato e delle esigenze del progetto al momento dell'inizio del tirocinio. I tirocinanti potranno svolgere parte del lavoro autonomamente con il proprio computer e saranno tenuti a un resoconto periodico della loro attività, nell'ottica di un dialogo costante tra tirocinante e tutor. Il progetto fornirà i registratori e i supporti di memoria necessari alla realizzazione delle registrazioni.

3. Risvolti applicativi del tirocinio

Il tirocinio permetterà di acquisire competenze di carattere altamente interdisciplinare e applicativo, che potranno essere utilmente impiegate nei seguenti ambiti lavorativi:

- Traduzione per il doppiaggio e sottotitolatura
- Realizzazione di sondaggi e interviste per agenzie private
- Utilizzo di dati linguistici a scopo commerciale (*data mining, sentiment analysis, …*)
- Trattamento automatico di dati linguistici
- Costruzione e gestione di database linguistici
- Conoscenza approfondita e utilizzo di *software* nell'ambito delle *digital humanities*

DIPARTIMENTO DI LINGUE, LETTERATURE E CULTURE MODERNE
ALMA MATER STUDIORUM • UNIVERSITÀ DI BOLOGNA

VIA CARTOLERIA, 5 – 40124 BOLOGNA – ITALIA

4. **Requisiti di ingresso e referenti per il tirocinio**

Gli studenti dovranno aver già sostenuto e superato almeno un esame di Linguistica generale.

Il referente per il tirocinio sarà la Prof.ssa Caterina Mauri (caterina.mauri@unibo.it).

Afferiscono al progetto anche la Dott.ssa Silvia Ballarè (silvia.ballare@unito.it), la Dott.ssa Eleonora Zucchini (eleonora.zucchini@unibo.it), la Dott.ssa Ludovica Pannitto (ludovica.pannitto@unibo.it), la Prof.ssa Claudia Borghetti (claudia.borghetti@unibo.it). Eventuali ulteriori collaborazioni saranno possibili su richiesta.

La coordinatrice del progetto,

Caterina Mauri
Prof.ssa associata di Linguistica Generale (L-LIN/01)
Università di Bologna - Dipartimento di Lingue, Letterature e Culture moderne
Email: caterina.mauri@unibo.it
