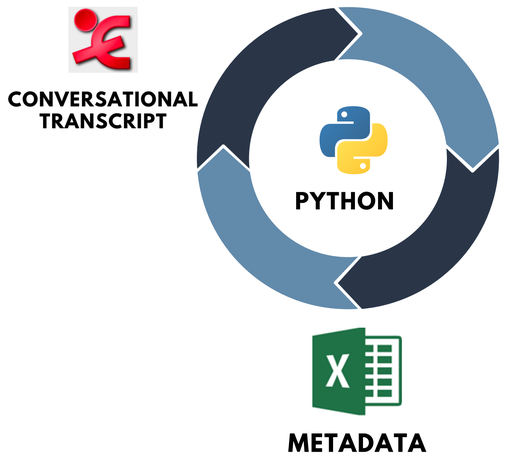
DiverSITa

A NEW CORPUS FOR THE DOCUMENTATION OF L2 ITALIAN

Eugenio Goria (Università di Torino)
Caterina Mauri (Università di Bologna)

KIParla is the most recent corpus of spoken Italian that is made publicly available.



NoSketch Engine

1.3 million words
more info @
www.kiparla.it

KIParla is a modular corpus where each module can be queried autonomously

- **KIP:** interaction at the University in Torino and Bologna (conversation, interviews, office hours, lessons, exams). Focus on register variation.
- **KIPasti:** dinner-table conversation involving families with different origins and social background. Focus on social variation.
- **ParlaTo:** Interviews collected in Torino. Focus on social variation.

The DiverSITa project aims to further expand the KIParla corpus by adding new data from L2ers and plurilingual speakers of Italian.

- To provide a better representation of plurilingualism
- To analyse the linguistic outcomes of recent migration flows
- To work towards recognition and valorisation of the repertoires of plurilingual speakers.

To account for L2 varieties of Italian and include variation in:

- educational backgrounds (school programmes, amateur courses, untutored learning),
- family language policies
- types of attainment.

- **ParlaBo:** Interviews collected in Bologna. Focus on social variation.
- **ParlaBz:** dinner-table conversations collected in Bolzano/Bozen
- **ParlaNa:** Interviews collected in Napoli. Focus on social variation.

The same collection methods are used for:

- Speakers of national origin
- Heritage speakers
- L2ers

The data collection so far

Community	Interview	Conversation	Tot
China BO	9:19:00	2:13:39	**11:32:39**
Ukraine BO	9:46:27	1:29:03	**11:15:30**
Morocco BO	8:32:34	1:35:54	**10:08:28**
Bangladesh BO	9:58:16	0	**9:58:16**
Peru To	06:28:15	00:47:18	**07:15:33**
Morocco TO	06:34:43	00:09:17	**06:44:00**
China TO	01:51:38	0	**01:51:38**
Romania TO	10:03:10	0	**10:03:10**
Tot	**37:23:40**	**5:18:36**	**68:49:14**

Expected: ca. 128 hrs, 1,250,000 tokens

Principles for the transcription of L2 varieties

- form-to function approach in order to avoid comparative fallacy (Rastelli 2009, Andorno 2009). Phenomena beyond the word level (e.g. agreement) are not tagged.
- Forms that are phonetically non standard but do not affect word morphology are normalised based on standard orthography.
- Inexistent forms are marked with a $ sign.
- Words in languages other than Italian are marked with a # sign.
- Forms that are phonetically ambiguous are normalised in writing, but a ( ) sign is used to mark the transcriber's best guess.

```
PST001  alla @chissia (.) @chissia #se #dice? #iglesia?
INT001  è chiesa chiesa

- at the chissia is chissia how you say 'church'?
- it's 'chiesa'
```

```
PST001  $arrivamo: in italia: con mio figlio,
INT001  mh mh,
PST001  e mio marito,
PST001  por #por #tema #de::la malattia al
        cuore di mio figlio

we arrive in Italy with my son, because of my son's heart disease
```

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA