Finanziato dall'Unione europea
NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA
ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

KIParla corpus of spoken Italian: design, usage and future challenges

Université de Neuchâtel

22/04/2024

Caterina Mauri (University of Bologna)

caterina.mauri@unibo.it

Overview

1. Background

2. Corpus design and implementation: goals, problems and solutions

- Modularity and incrementality
- Corpus design
- Data collection: privacy, registration and management
- Transcription of data
- Data publishing: NoSketch Engine

3. Using the corpus

- *Come cosa*: the emergence of a new construction

4. Next steps and future challenges

Overview

1. Background

2. Corpus design and implementation: goals, problems and solutions

- Modularity and incrementality
- Corpus design
- Data collection: privacy, registration and management
- Transcription of data
- Data publishing: NoSketch Engine

3. Using the corpus

- *Come cosa*: the emergence of a new construction

4. Next steps and future challenges

The coordinators of the KIParla corpus

Caterina Mauri (University of Bologna)

- Linguistic typology;
- Language change;
- Intra- and interlingual variation;
- Semantics and pragmatics.

Silvia Ballarè (University of Bologna)

- Sociolinguistics;
- Intra- and interlingual variation;
- Contact between Italian and dialects.

Eugene Goria (University of Turin)

- Sociolinguistics;
- Heritage languages;
- Language contact.

Massimo Cerruti (University of Turin)

- Sociolinguistics;
- Intra- and interlingual variation;
- Contact between Italian and dialects.

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Brief (funding) history

2015-2019. Funding: Italian Ministry of University and Research

- SIR Program, Project No. RBSI14IIG0 **'LEADhoC - The linguistic expression of ad hoc categories'**

PI Caterina Mauri **-** www.leadhoc.org

– 2016: construction and communication of **ad hoc categories** in interactional spoken discourse.

✓ Design of a new corpus of spoken Italian

2019-2023. No funding – but the project kept growing… self-sustainability by means of university internship

2023-2025. Funding: Italian Ministry of University and Research, PNRR 2022*

- PRIN 2022 PNRR Program, Project No. P2022RFR8T **'DiverSIta – Diversity in Spoken Italian'**

PI Caterina Mauri - https://site.unibo.it/divers-ita/en

– Expansion on the KIParla corpus, including data of non-native speakers

** The research leading to these results has received funding from Project "DiverSIta-Diversity in spoken Italian", prot. P2022RFR8T, CUP J53D23017320001, funded by EU in NextGenerationEU plan through the Italian "Bando Prin 2022 - D.D. 1409 del 14-09-2022"*

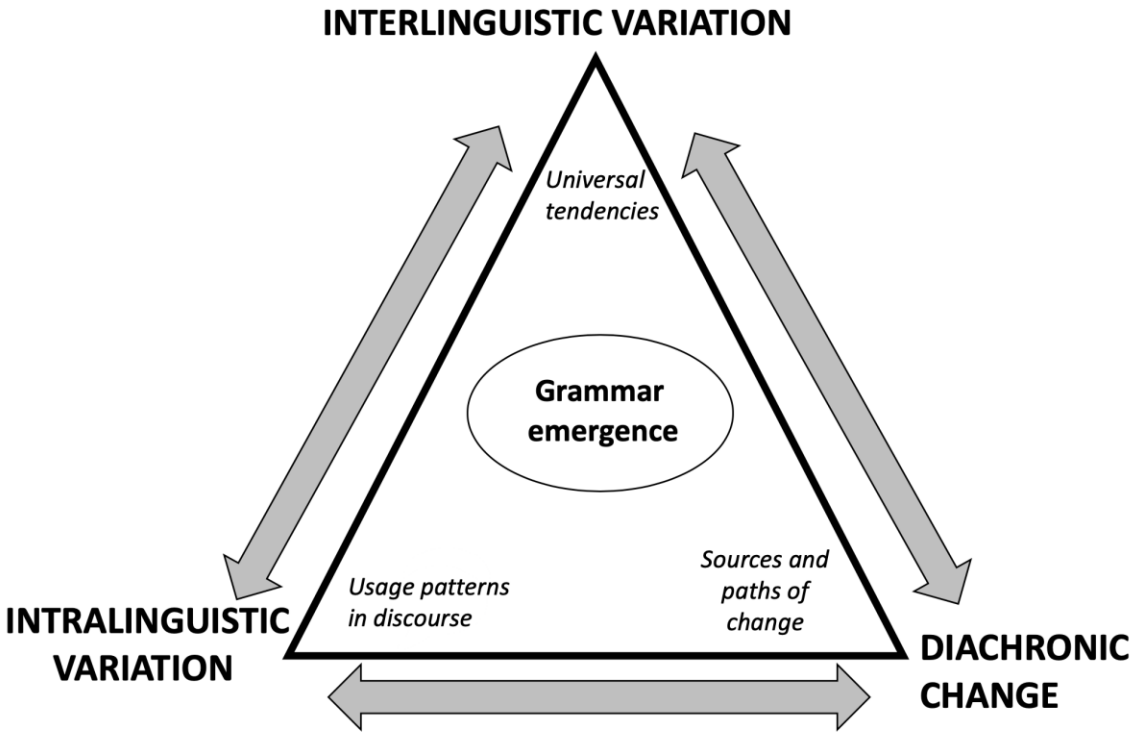
Why a corpus of Spoken Italian?

Integrated 3D methodology:
Discourse, Diachrony, Diversity

The larger picture…

- ✓ Functional-typological literature (Givón 2001; Croft 2003; Cristofaro 2012; Bybee 2008; Bybee and Beckener 2015, among others): close (inter)connection between cross-linguistic variation, discourse variation and diachronic change.

- ➢ Triangular view of language which calls for an **integrated 3D research methodology**!



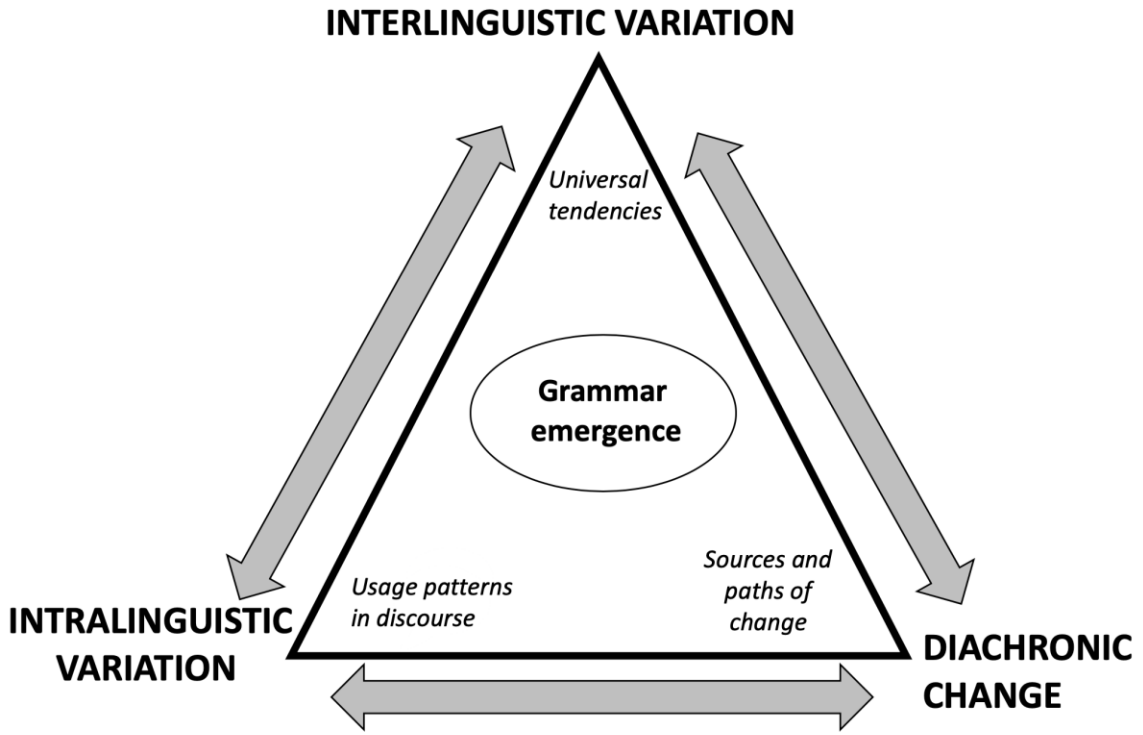
Adapted from Ballarè, Goria and Mauri 2022
Mauri and Masini 2022

Why a corpus of Spoken Italian?

Bybee and Beckener (2015: 197) nicely summarize the *rationale* underlying a 3D methodology as follows:

> "... we look for "universals" not by proposing strict definitions and properties all languages are expected to have, but instead by studying dynamic aspects of language **at the level of use, at the level of language-specific structures, and at the cross-linguistic level** (Croft 2001; Bybee 2010). Both similarities and differences will be instructive and help us to understand the on-line processes that create language structure as well as the biological, cognitive, and social factors that determine the outcome of change."

Integrated 3D methodology:
Discourse, Diachrony, Diversity



Ballarè, Goria and Mauri 2022,
Mauri and Masini 2022

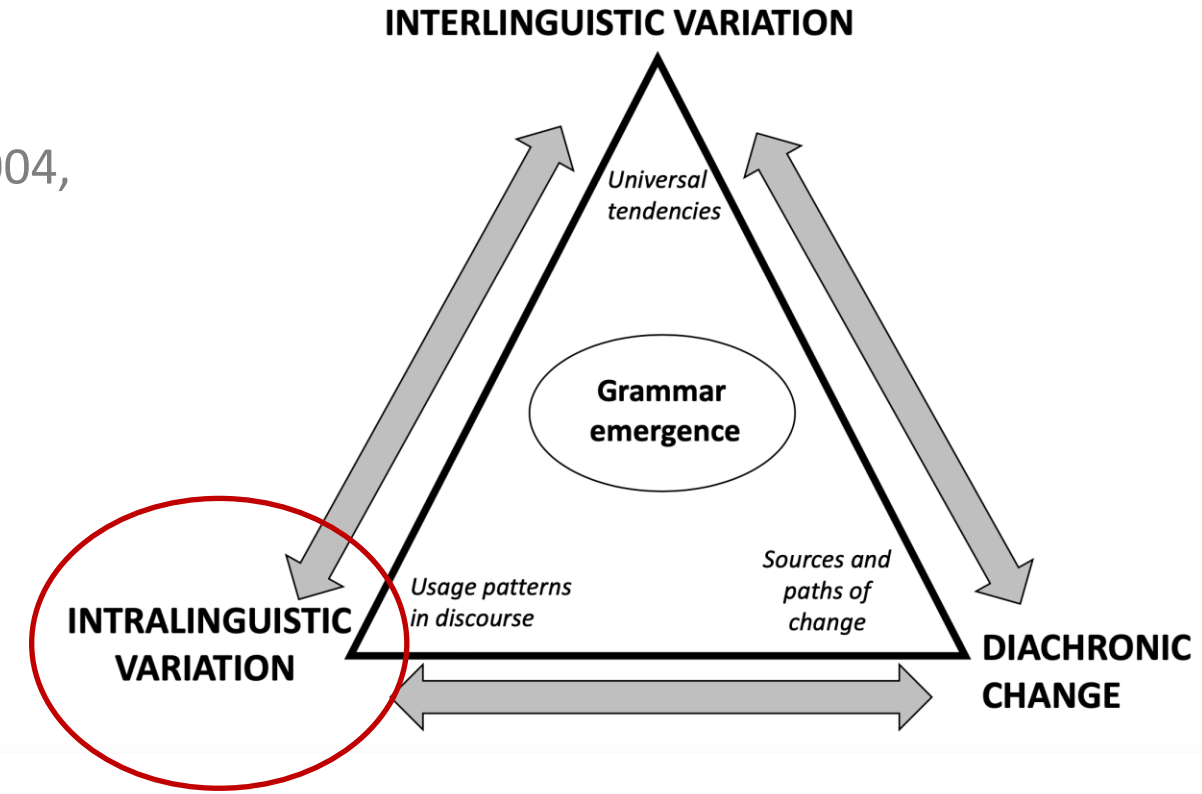
Why a corpus of Spoken Italian?

Integrated 3D methodology:
Discourse, Diachrony, Diversity

- ✓ **Patterns of use in discourse** are the *locus* where grammatical structures emerge out of repeated individual usage events, as a result of parsing preferences and efficiency choices (cf. Hawkins 2004, Bybee 2008, Voghera 2017).

- ✓ As a consequence, in the great distributional and functional variation that can be observed in discourse, we often see the reflections of successive diachronic stages in the **emergence of new grammatical constructions**

- ➢ The observation of spoken interactions allows to identify the **initial stages and the triggering factors of language change**



Ballarè, Goria and Mauri 2022,
Mauri and Masini 2022

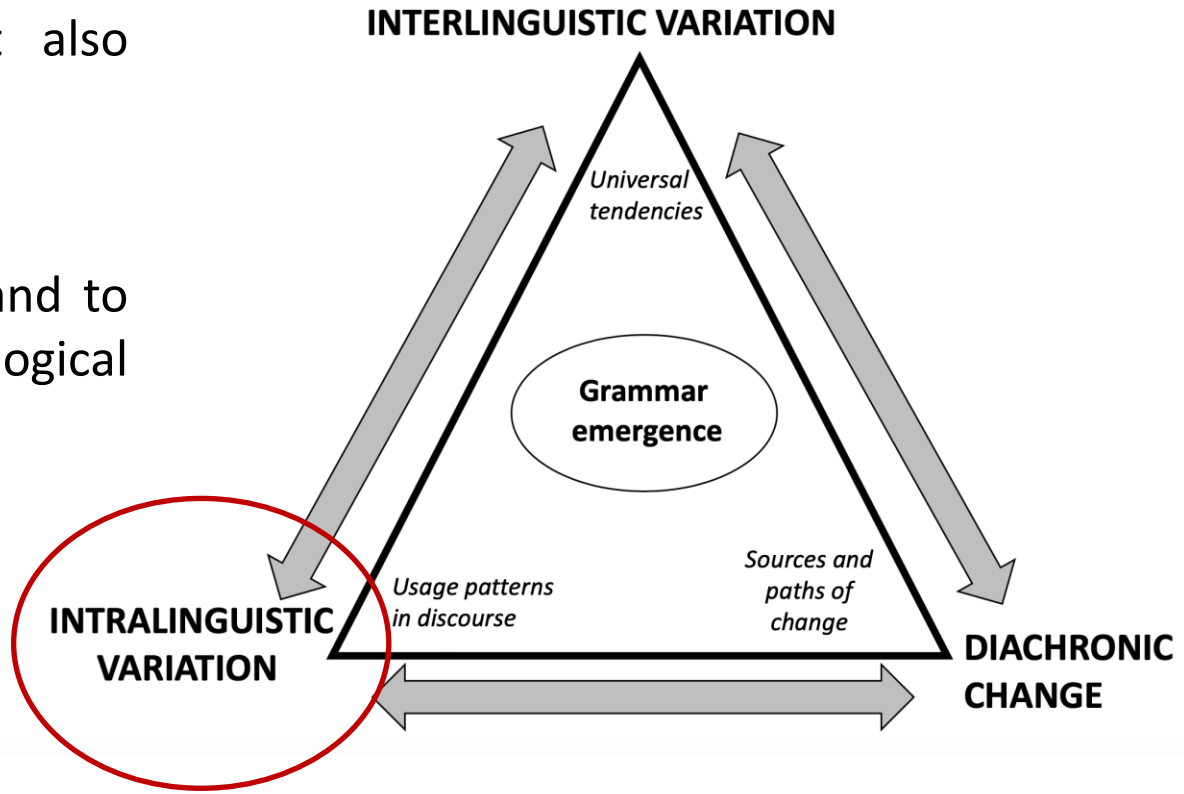
Why a corpus of Spoken Italian?

How grammar emerges from usage, and in particular from spontaneous speech, is not just fascinating but also challenging!

- inherent complexity of spoken language
- evidence from spoken data is harder to collect and to analyze, in both intra-linguistic and typological perspective.

What resources did we have in 2016 for the analysis of Spoken Italian?

Integrated 3D methodology:
Discourse, Diachrony, Diversity



Ballarè, Goria and Mauri 2022,
Mauri and Masini 2022

Corpora of spoken Italian: the background

✓ Various speech corpora, constructed for specific research purposes, with different methodologies (e.g. map task, interviews, ...) by individual scholars or research groups

... some of which have never been published!

✓ Some medium- and large-sized corpora, representative of internal variation in spoken Italian, have been made publicly accessible:

VoLIP Corpus	Data from the 1990s, 60 hrs. Transcripts and audio files, no KWIC search	https://www.volip.it	(Voghera et al. 2014)
LABLITA Corpus	Data collected between 1970s and 1990s, 200 hrs.	Not accessible	(Cresti and Moneglia 2005)
CLIPS Corpus	15 collection points, 100 hrs. Transcripts and audio files, no KWIC search	http://www.clips.unina.it/it/corpus.jsp (!)	(Sobrero and Tempesta 2007)

Corpora of spoken Italian: the background

! **Accessibility and maintenance** issues

- Different search interfaces: CQP or ad hoc interfaces;
- Corpora not easily accessible (on antiquated media, or on inactive servers)-the only one fully accessible is VoLIP;
- KWIC search not always available

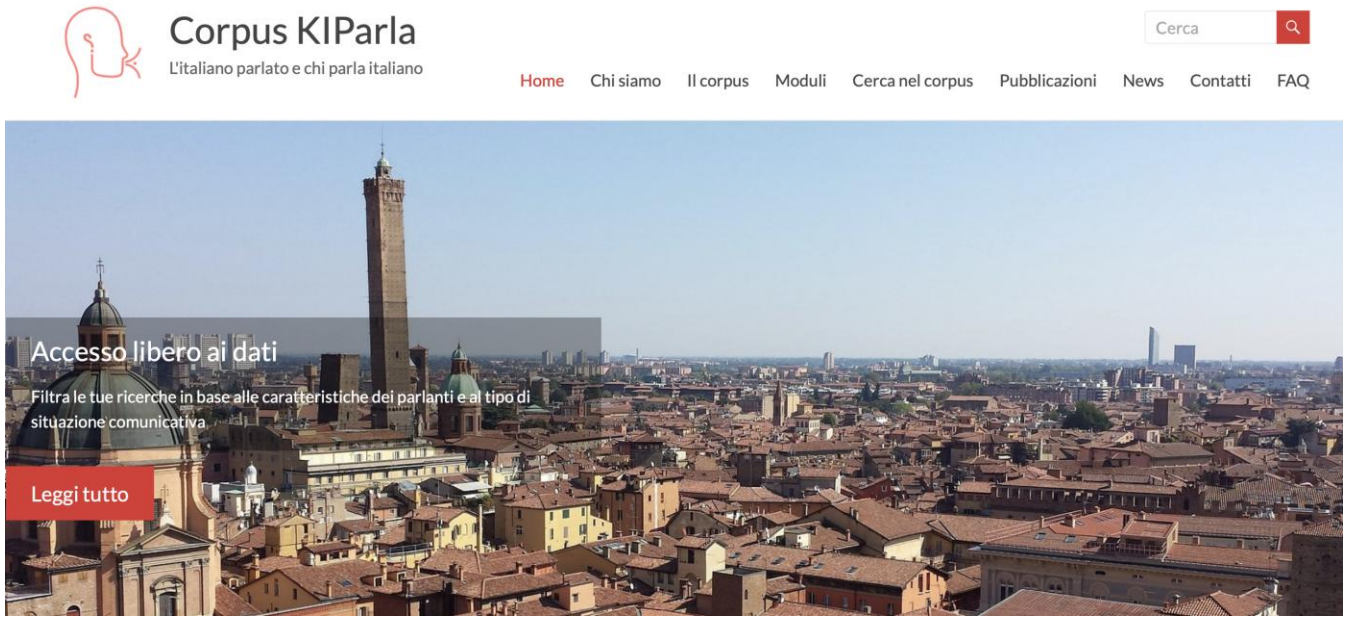
= Issues of **comparability and analysis**

- Data with little integration, having been collected on the basis of different parameters and scientific needs;
- Audio not always accessible;
- There is no direct connection between transcription given audio-except in VoLIP!

? Few **metadata on speakers**

- Difficult to consider diatopic variation: collection point vs. origin of speakers;
- Almost impossible to investigate diastratic variation;
- Relationship between interlocutors not known-except in VoLIP!

Design of the KIParla ~~corpus~~ enterprise!

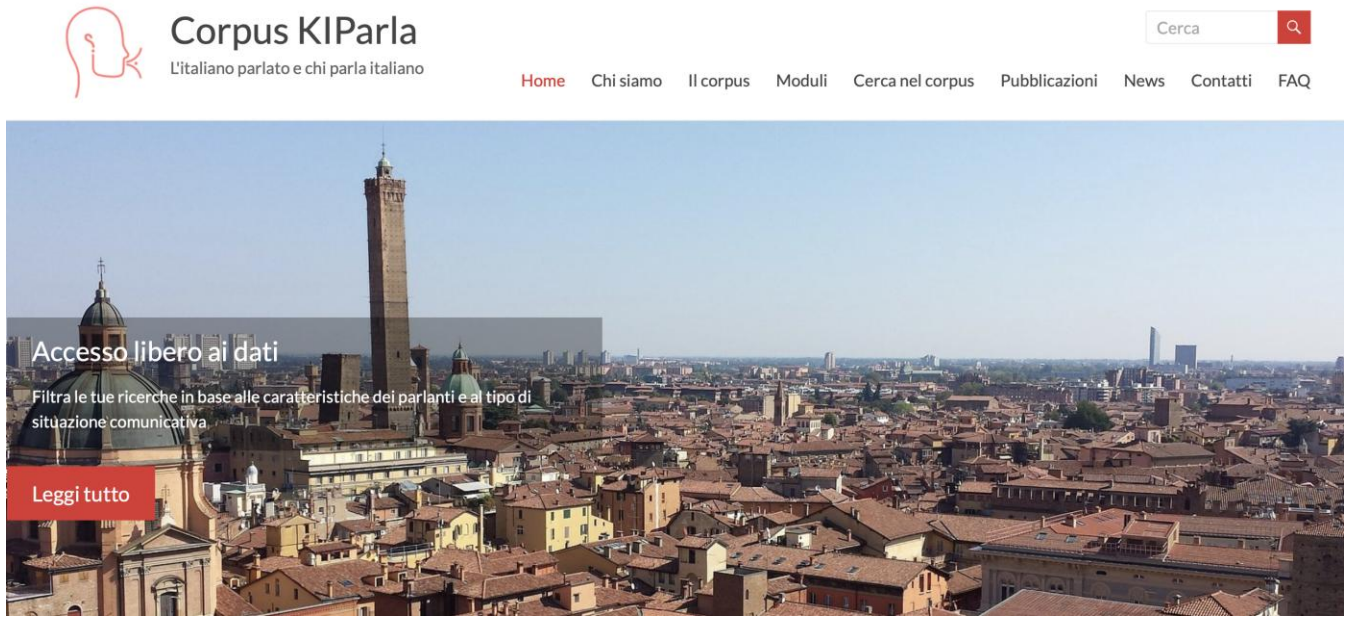


www.kiparla.it

The KIParla corpus aims to offer:

- ✓ **freely accessible** corpus
  transcripts aligned with audio files
- ✓ **transparent metadata** system
  both wrt the communicative situations and the speakers involved
- ✓ **search interface** based on an **international standard**
  NoSketchEngine, offering advanced search functions: KWIC, frequency lists, etc. (see §3)
- ✓ **modular, incremental, replicable infrastructure**
  allowing for the expansion of the corpus over time through a modular structure

Design of the KIParla corpus



www.kiparla.it

The KIParla corpus aims to offer:

- ✓ **freely accessible** corpus
  transcripts aligned with audio files
- ✓ **transparent metadata** system
  both wrt the communicative situations and the speakers involved
- ✓ **search interface** based on an **international standard**
  NoSketchEngine, offering advanced search functions: KWIC, frequency lists, etc. (see §3)
- ✓ **modular, incremental, replicable infrastructure**
  allowing for the expansion of the corpus over time through a modular structure

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Overview

1. Background

2. Corpus design and implementation: goals, problems and solutions

- Modularity and incrementality
- Corpus design
- Data collection: privacy, registration and management
- Transcription of data
- Data publishing: NoSketch Engine

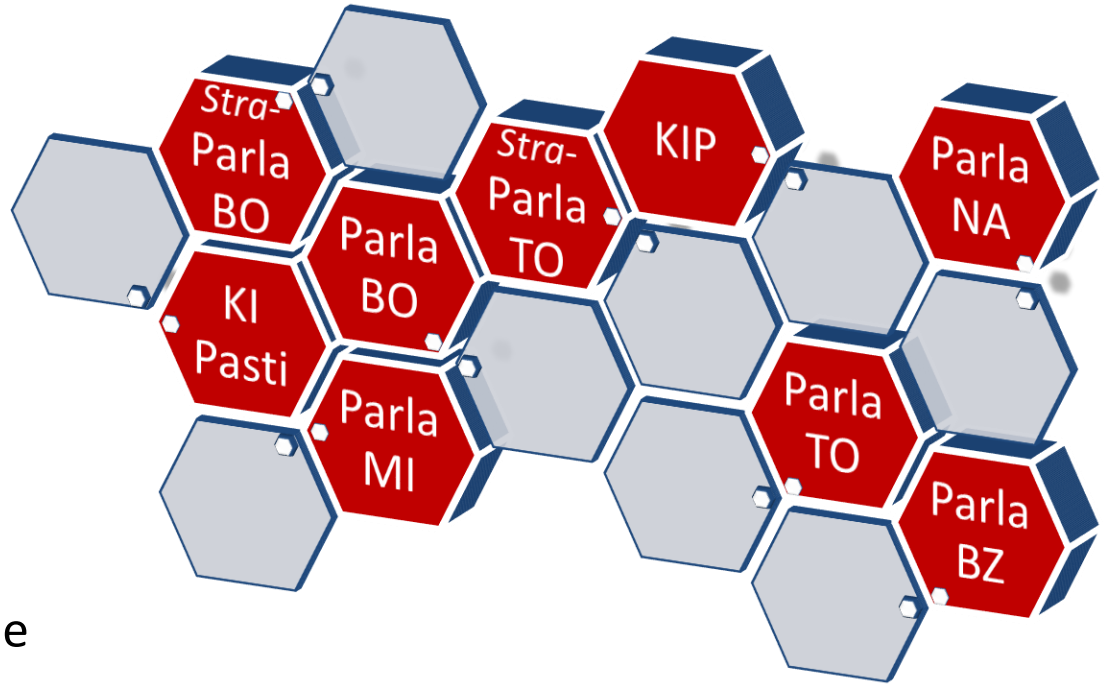
3. Using the corpus

- *Come cosa*: the emergence of a new construction

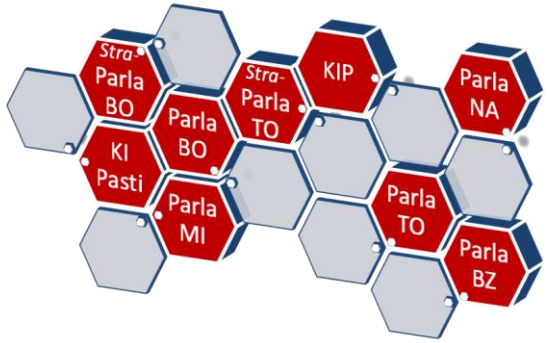
4. Next steps and future challenges

Modularity and incrementality

- **Modularity:** internal division into independent modules
  - ✓ possibility of consulting each module separately or all modules together;
- **Incrementality**: possibility to add new modules over time



Modularity and incrementality

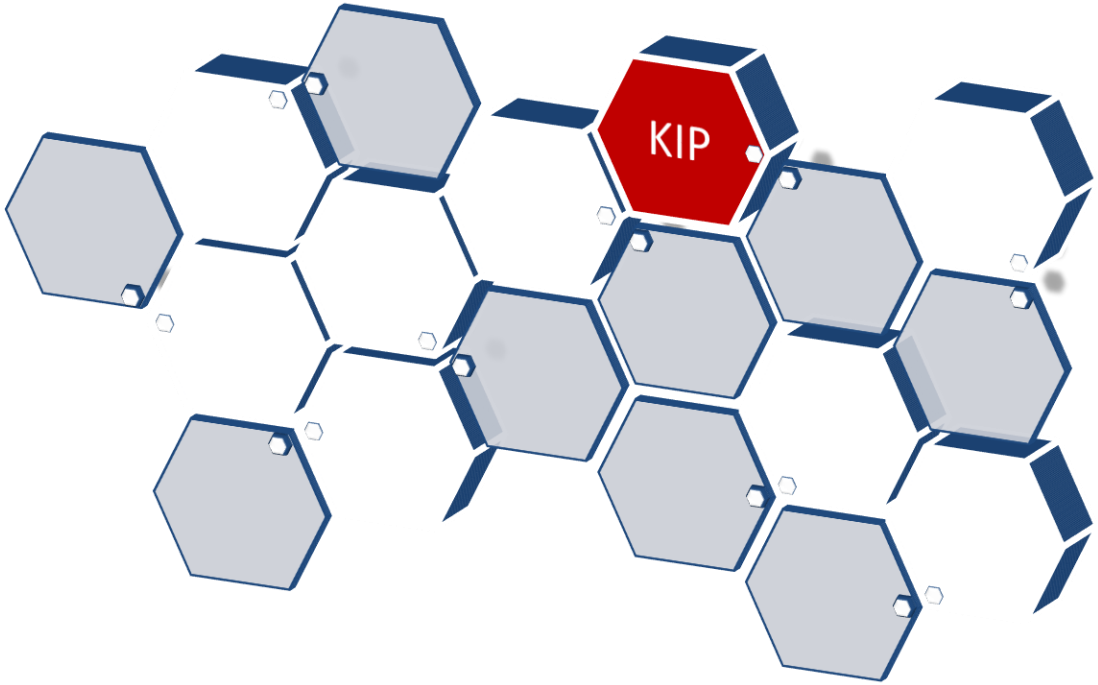


Modules are small/medium sized corpora

- ✓ built for different, often complementary purposes
- ✓ with data from different geographic areas, different types of interaction and different types of communities
- ➢ while maintaining **substantial underlying comparability in structure and accessibility**.

- ➢ **The comparability and at the same time the specificity of each module are what can make the KIParla corpus representative of spoken Italian**: the more modules that are added, through the collaboration of different universities, the more dimensions of variation can be explored.

Corpus design: the KIP module



Step one: **KIP module**

- language spoken in a university setting

→ Observation of register and regional variation.

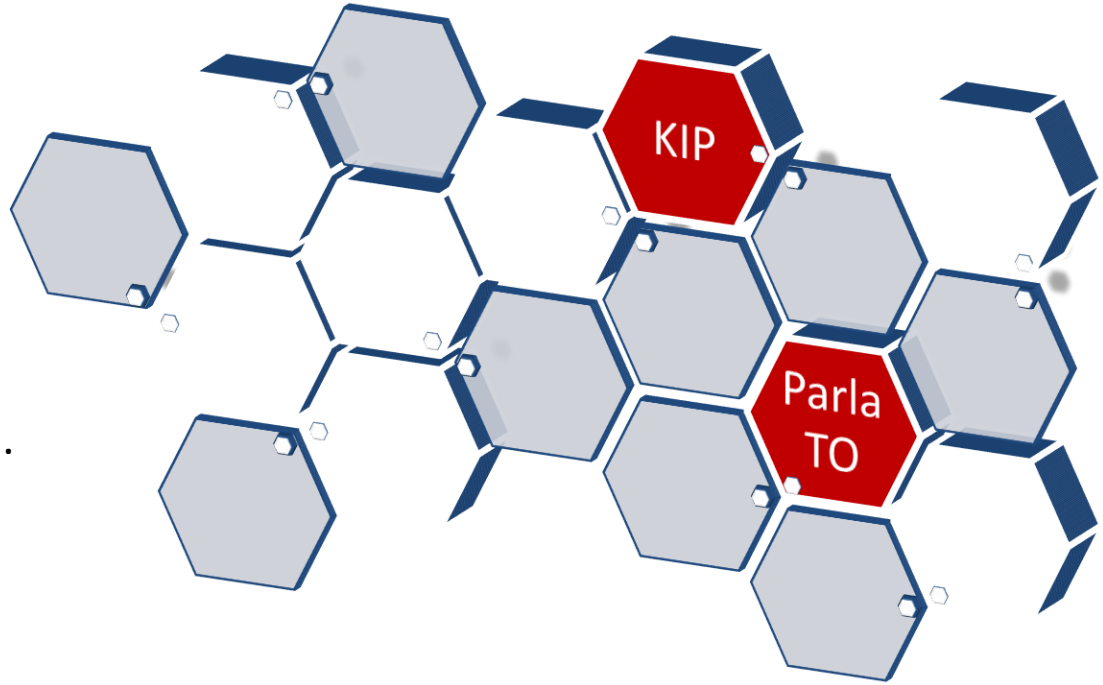
- ✓ Data collected in Bologna and Turin (2016-2019)
- ✓ University students and professors: **educated speakers**
- ✓ Different communicative contexts
- ✓ **TOT: ca. 70 hrs, 660,000 tokens**

	Relationship between participants	Moderator	Topic
Free conversation	Symmetrical	Absent	Free
Semi-structured interview	Symmetrical	Present	Fixed
Student reception	Asymmetrical	Absent	Free
Examinations	Asymmetrical	Absent	Free
Lessons	Asymmetrical	Present	Fixed

PUBLISHED

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Corpus design: the ParlaTO module



Second step: **ParlaTO module**

- language spoken in the metropolitan city of Turin

→ Observation of the vernacular, diastratic and regional variation.

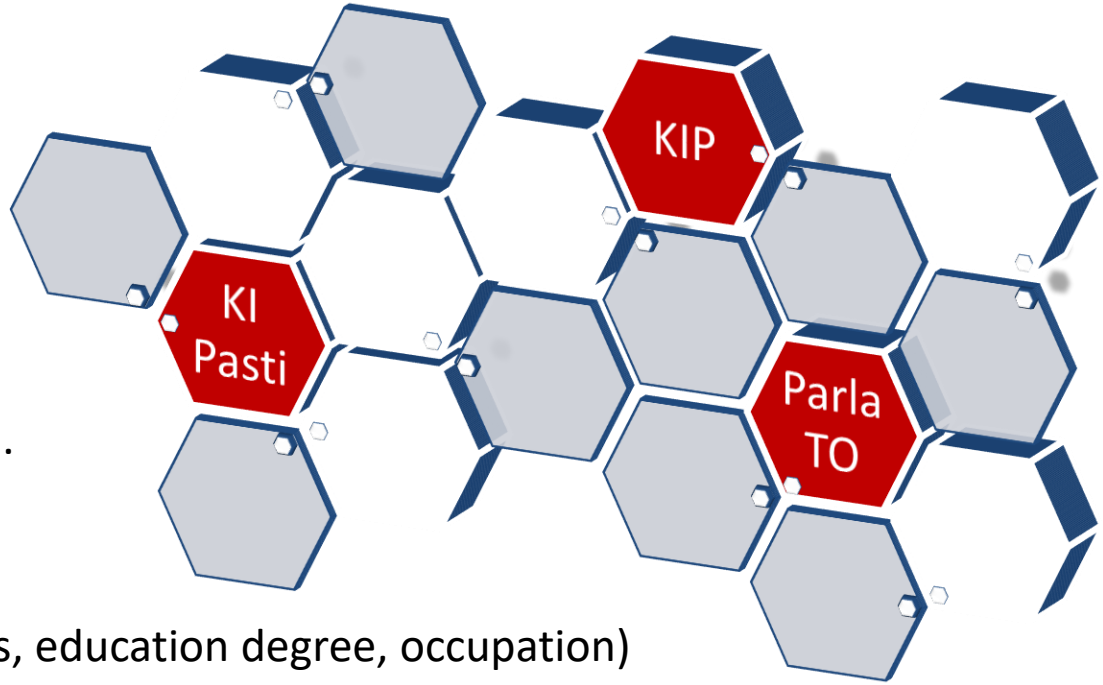
- ✓ Data collected in Turin (2019)
- ✓ Speakers with **different social characterization** (ages, education degree, occupation)
- ✓ One context: the semi-structured interview
- ✓ **TOT: ca. 50 hrs, 560,000 tokens**

	Age groups
Young	18 ≤ x ≤ 30 years old
Adults	30 < x ≤ 60 years old
Seniors	60 < x ≤ 89 years old

PUBLISHED

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Corpus design: the KIPasti module



Third step: **KIPasti module**

- Language spoken at 63 dinner-table conversations

→ Observation of the vernacular, diastratic and regional variation.

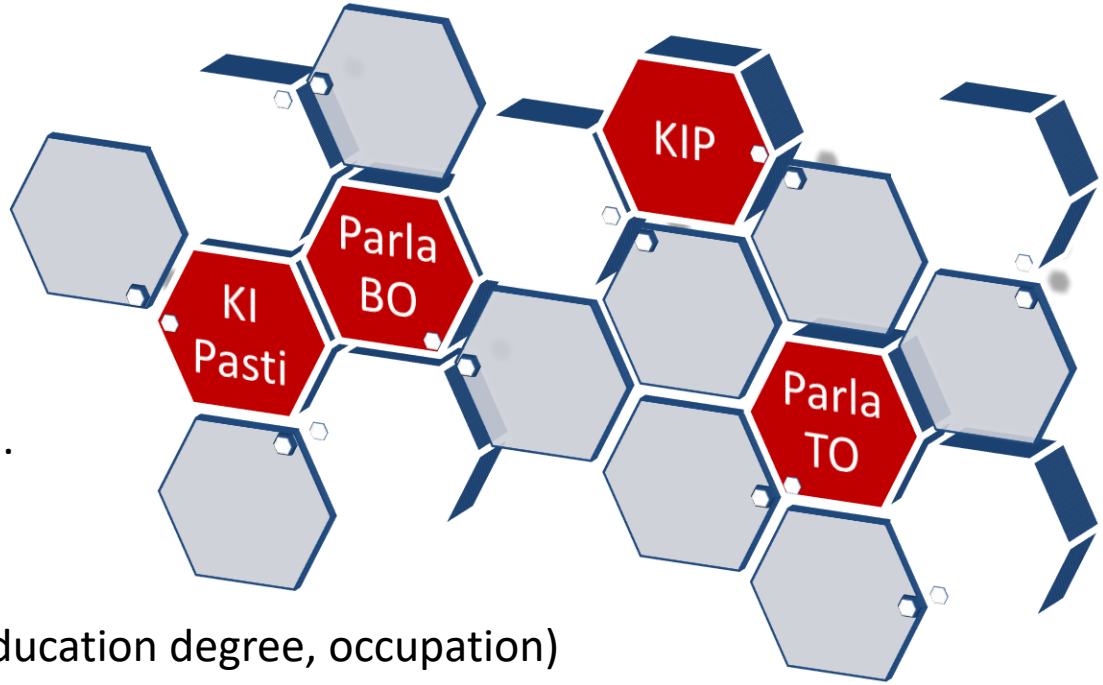
- ✓ Data collected in 13 regions of Italy (2020-2024)
- ✓ 147 Speakers with different social characterization (ages, education degree, occupation)
- ✓ One context: **dinner-table conversations**, thus shared background, informal and friendly register
- ✓ **TOT: ca. 42 hrs, 487,000 tokens**

	Data collection
Northern Italy	46%
Central Italy	16%
Southern Italy	38%

PUBLISHED

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Corpus design: the ParlaBO module



Fourth step: **ParlaBO module**

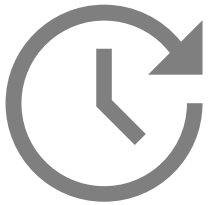
- Language spoken in the metropolitan city of Bologna

→ Observation of the vernacular, diastratic and regional variation.

- ✓ Data collected in Bologna (2022-2024)
- ✓ Speakers with **different social characterization** (ages, education degree, occupation)
- ✓ One context: the semi-structured interview
- ✓ **TOT: ca. 66 hrs, 650,000 tokens**

	Data collection
18 ≤ x ≤ 30 years old	30%
30 < x ≤ 60 years old	36%
60 < x ≤ 89 years old	34%

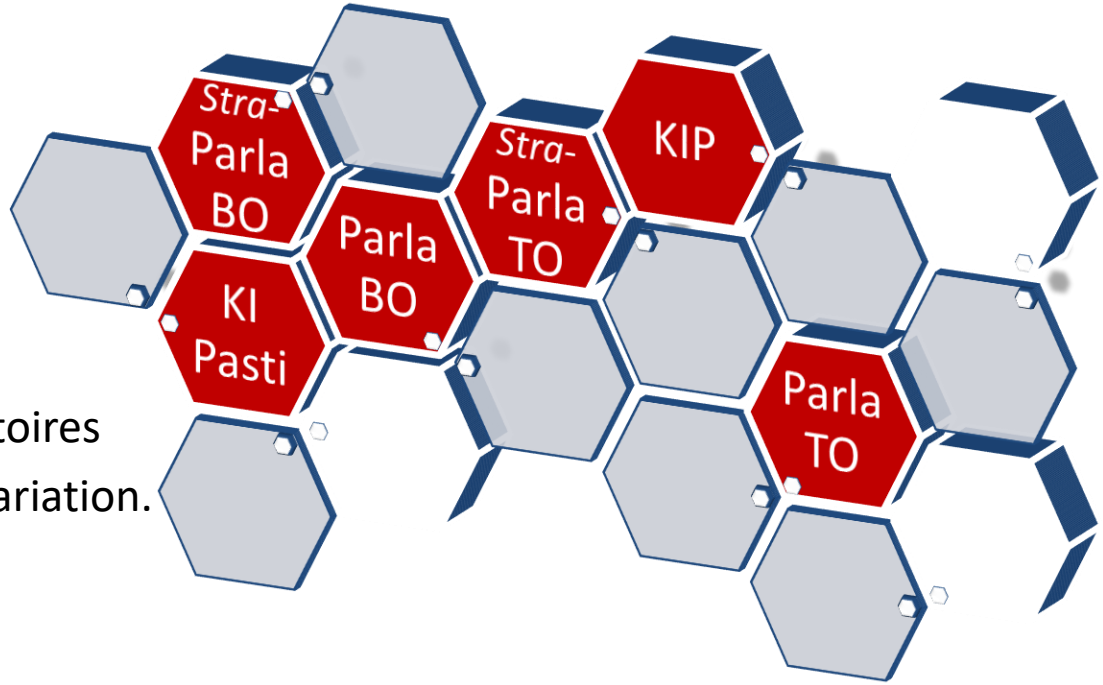
- ✓ Data collection completed
- ➢ At the moment, we are transcribing (70%)



TO BE PUBLISHED
June 2024

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Corpus design: Stra-Parla BO/TO modules



Next steps: **Stra-ParlaBO and Stra-ParlaTO modules**

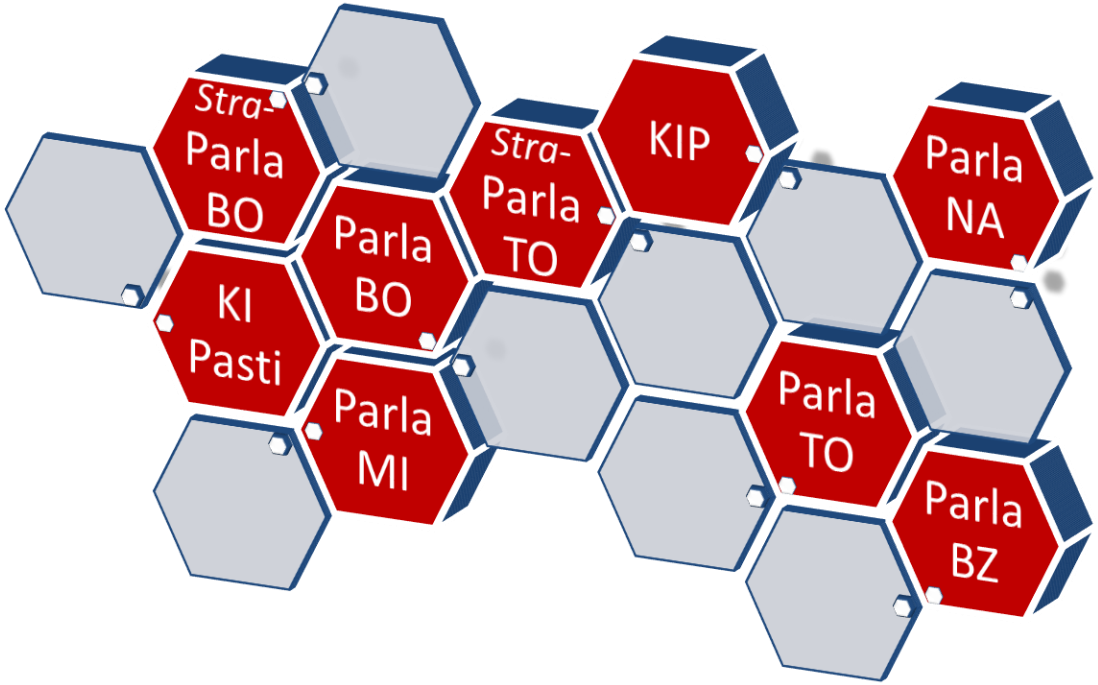
- Italian spoken within communities of foreign-origin speakers with complex multilingual repertoires

→ Observation of the learners' varieties, diastratic and regional variation.

- ✓ Data collected in Bologna and Turin (2024 - ongoing)
- ✓ Speakers with multilingual repertoire, different social characterization and origin
- ✓ Two contexts: semi-structured interview, dinner-table conversation
- ✓ **EXPECTED: ca. 128 hrs, 1,250,000 tokens**

Data collection in progress!

Corpus design: ongoing collaborations



Next steps: **ParlaBZ, PArlaNA, ParlaMI modules**

- Italian spoken in Bozen, Naples, Milan
- Universities of Bolzano, Naples, Milan Bicocca

→ Observation of the vernacular, diastratic and regional variation

- ✓ Data collected in Bolzano, Naples and Milan (ongoing)
- ✓ Speakers with different social characterization and origin

Data collection in progress thanks to the collaboration with colleagues of different universities!

Collection and transcription: *it takes a village!*

2018 - today: more than 100 students (from the universities of Bologna and Turin) participated in the construction of the KIParla corpus.

- Internships, three-year theses, master's theses, ...

Training and supervision

Frequent updates and weekly meetings

Organization and coordination

Data collection: the legal side

GDPR - Regolamento 2016/679

- Who is responsible for the data?
- What metadata will be stored? Will they be aggregated?
- Where will the data be stored?
- Who has access to the data?
- Do the contents of the records present sensitive information (first and last names, addresses, ...)?

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

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Transcription and anonymization of data

Most of the data were transcribed by students of Bologna universities by means of ELAN.

- Training;
- Periodic monitoring.

A simplification of the system proposed by Jefferson (2004) was employed to account for conversational aspects (overlaps, nonverbal behaviors, ...).

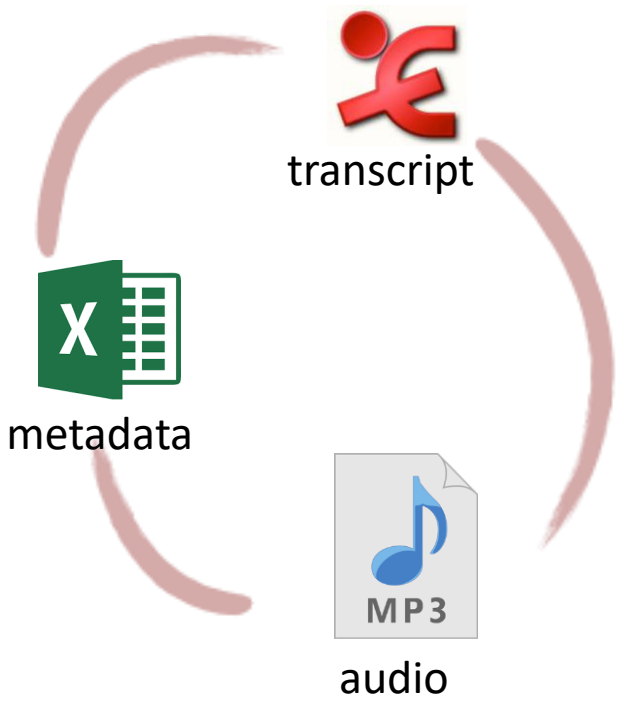
,	Rising intonation	<ciao>	Pronunciation (slower)
.	Descending intonation	[hello]	Overlaps between speakers
:	Prolonged sound	(hello)	Text difficult to understand
(.)	Short break	xxx	Unintelligible text
>ciao<	Pronunciation (faster)	((laughs))	Nonverbal behavior

Anonymization:

- Deletion of sensitive data from the transcript and audio track.

Publication of data

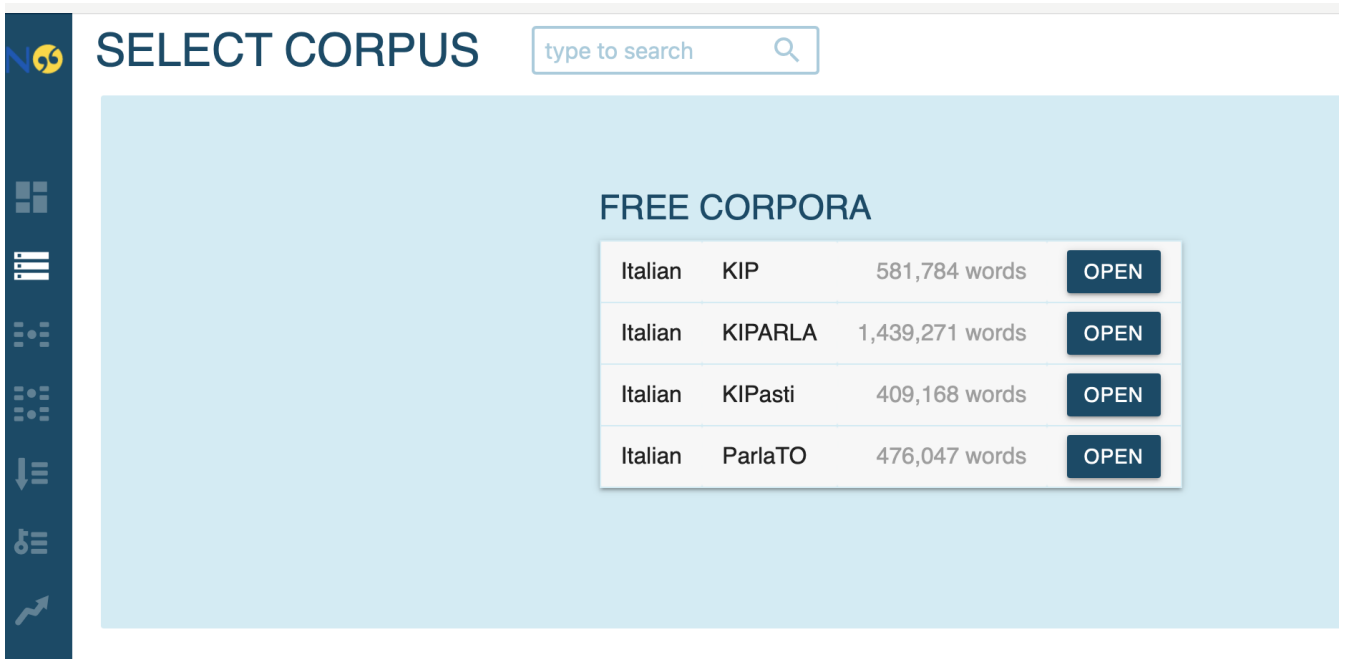
- XML
- Txt (spelling version)
- Txt (conversational version)
- Occurrence/metadata association
- Alignment with audio



scripts in python

- International standard
- Advanced functionality
- Open source

NoSketch Engine



KIParla corpus: published modules

- ✓ KIP
- ✓ ParlaTO
- ✓ KIPasti

KIParla Corpus (22/04/24)	
Hours	152:50:25
Informants	489
Tokens	1,439,271

Accessibility:

Transcriptions (spelling and conversational):

- **freely** accessible through NoSketch Engine, no registration required;

Audio:

- **Freely** accessible **upon registration**, you can listen to the audio tracks online (from NoSketch Engine);

https://kiparla.lingue.unibo.it/kiparla/crystal/#open

KIParla corpus: upcoming

- ✓ ParlaBO
- ✓ ParlaBZ
- ✓ Stra-ParlaBO
- ✓ Stra-ParlaTO

KIParla Corpus (30/11/25)	
Expected Hours	343:00:00
Expected Informants	900
Expected TOT Tokens	3,300,000

Accessibility:

Transcriptions (spelling and conversational):

- **freely** accessible through NoSketch Engine, no registration required;

Audio:

- **Freely** accessible **upon registration**, you can listen to the audio tracks online (from NoSketch Engine);

Guidelines for a Modular, Incremental, Replicable resource (MIR)

Our goal is not only to build a Modular, Incremental, and Replicable resource (MIR resource), but also to make the whole **process fully accessible and replicable** also for other languages in and outside Europe.

- All the methodological choices are being documented at every stage

- We will publish a set of guidelines for the development of Modular, Incremental and Replicable (MIR) resources to document orality, in compliance with the FAIR principles (Wilkinson et al. 2016).

Overview

1. Background

2. Corpus design and implementation: goals, problems and solutions

- Modularity and incrementality
- Corpus design
- Data collection: privacy, registration and management
- Transcription of data
- Data publishing: NoSketch Engine

3. Using the corpus

- *Come cosa*: the emergence of a new construction

4. Next steps and future challenges

Using the corpus / 1

Types of research:

- ✓ Query types
- ✓ Context
- ✓ Text types

Ci sta, lit. 'there it fits' > response particle, agreement marker (Battaglia & Mauri, in prep.)

Per cui From oblique relative clause 'for which' > to conclusive connective (Ballarè, Mauri, in prep.)

Bella come cosa! Topicalizing and reinforcing uses of [Pred [*come NP*]] (Ballarè, Goria, Mauri 2022)

Other features:

- ✓ Frequency lists (Word lists, frequency lists for each metadata)
- ✓ Corpus info

www.kiparla.it

https://kiparla.lingue.unibo.it/kiparla/crystal/#concordance?corpname=KIPARLA

Using the corpus / 2

KWIC:

- ✓ Save concordance (txt, csv, xml)
- ✓ Frequency lists (Word lists, frequency lists for each metadata)
- ✓ View options: different metadata can be selected to be displayed and exported along with occurrences
- ✓ Sort
- ✓ Random sample
- ✓ Additional filters
- ✓ Frequency analysis of occurrences (in relation to different metadata)
- ✓ Collocation candidates

www.kiparla.it

https://kiparla.lingue.unibo.it/kiparla/crystal/#concordance?corpname=KIPARLA

Using the corpus / 3

Metadata and links:

- ✓ Click on KWIC > context expansion
- ✓ Click on the conversation code
  - ➢ All metadata related to the speaker and conversation
  - ➢ Link to the conversation in html format (orthographic)
  - ➢ Link to the conversation in html format (conversational)
  - ➢ Link to audio file aligned to specific occurrence (3 sec. earlier)

www.kiparla.it

https://kiparla.lingue.unibo.it/kiparla/crystal/#concordance?corpname=KIPARLA

Let's focus on [P_{red} [come N_{bare}]]: *a qualitative perspective*

Let’s consider these examples

(1) BO113: non è che devo per forza stare con uno che **[non mi piace_$_{Pred}$ [come atteggiamenti]]**
It's not that I have to be with someone that **[I don't like_$_{Pred}$ [in terms of attitudes]]**
(KIParla-KIP, BOD2014)

✓ The evaluation *non mi piace* ‘I don’t like’ does not concern the totality of the person in question, but should be referred to a **narrow domain**, that is, his or her *attitudes*

(2) BO014: di dublino mi e' piaciuto vabbe' il centro / of dublin I liked *vabbe* the center
eh **[[mi ricordava un po' bologna]$_{Pred}$** come / eh **[[it reminded me a little bit of Bologna]$_{Pred}$** as
TO999: mhmh / mhmh
BO014: **come architettura]** **/ as architecture**]
(KIParla-KIP, BOD2021)

✓ In (2) the statement *Dublin [...] reminded me a bit of Bologna* should be referred to the **domain** of *architecture* only.

Let’s consider these examples

In the construction observed in (1) and (2), *come* does not have similative or equative value, but has developed a function related to the **organization of information**, specifically N is the domain for which the predication is valid.

Let's consider these examples

In the construction observed in (1) and (2), *come* does not have similative or equative value, but has developed a function related to the **organization of information**, specifically N is the domain for which the predication is valid.

BUT

(3) TO038: […] sarebbe figo fare un posto che non esiste / it would be cool to make a place that doesn't exist
solo cocktail a chilometri zero / just zero-mile cocktails
non esiste […] / does not exist [...]
TO031: no minchia in realtà [[**è fighissima**]$_{Pred}$ **come cosa**] / actually [[**that's cool**]$_{Pred}$ **as a thing**]
non pensavo che tu volessi fare un po' il barista / I didn't think you wanted to be a barman

(KIParla-KIP, TOA3004)

…What domain???

Beyond functive *come*

- ***Functive*** uses : Creissels 2014, *role phrase* (Haspelmath/Buchholz 1998).

  A specific role or function of a verb arguments (e.g., *works **as a pizza maker***).

➢ The construction in (1) and (2) is similar to the functive usages but, unlike the latter, it does **not specify any role** but rather **delimits the domain and topic of the predication**.

➢ Moreover, N not only narrows the semantic domain over which the predication has value, but also behaves as **a displaced topic, which** the speaker decides to explicate, often as an **afterthought** (cf. anti-topic, Lambrecht 1981).

➢ Corpus-driven analysis based on the KIParla corpus

Emergent construction [P_{red} [come N_{bare}]]

- The variation observed in the data points to an emerging construction (cf. Hopper 1987, 2011; Auer & Pfänder 2011; Calaresu 2018) in spoken Italian, characterized by a **predication** preceded or followed by the phrase [*come N*], where the noun introduced by *come* is always bare, singular or plural, and cannot be accompanied by either determiners or modifiers

[P_{red} [*come* N_{bare}]] [[*come* N_{bare}] P_{red}]

Functive *come*:
Indicates the role
or function
of a verbal argument

Delimitative *come*:
Indicates the domain
for which the
predication holds true

Topical-anaphoric *come*:
Indicates or resumes
the discourse topic

Functive *COME*

(4)BO028: sì grazie / yes thank you

BO026: però [[**mi mandate anche una mail**]$_{Pred}$ **come promemoria**] / [[**you also send me an email**]$_{Pred}$ **as a reminder**]

BO028: certo / sure

(KIParla-KIP BOA1002)

- N indicates a specific **role** or **function** of one of the verb arguments

- ✓ TYPES OF N: often **animate** (48%, frequent lexical field of **professions**)
- ✓ PREDICATE TYPES: nonspecific, rarely evaluative (2%)
- ✓ CONTEXT: roles, functions, tools, interpretations

Delimitative COME

(5) pero' lowenbrau
[**[mi sembra più tedesco]$_{Pred}$ come nome**]

/though lowenbrau
/[[sounds more German to me]$_{Pred}$ as a name]

(6) con la cella frigorifera che [**[è molto più più grande]$_{Pred}$ come superficie**]
(KIParla - KIP BOA3021)

/with cold storage that [[is much larger]] $_{Pred}$ as a surface] (LIP MC10)

- N provides the **domain or scope** needed to understand the predication

- ✓ TYPES OF N: rarely animate (16%)
- ✓ PREDICATE TYPES: tendency for **evaluative predicates** (70%)
- ✓ RARELY PREPOSED to predication (16%, e.g., ***[come zona]*** *[piu' o meno e' limitrofa]$_{Pred}$*)

From delimitative COME to topical-anaphoric *COME*

(7) no fidati se lo conoscessi sarebbe divertente
perché [**[lui è proprio divertente]**$_{Pred}$ **come persona**]
(KIParla-KIP BOA3021)

/ no trust me if I met him it would be funny
/ because [**[he's just funny]**$_{Pred}$ **as a person**]

(8) non tanto per pe per il fatto che son tossici ma per
il fatto che stanno insieme […]
che che [**[non va bene]**$_{Pred}$ **come cosa**]
(KIParla-KIP BOA3021)

/ not so much for pe for the fact that they are drug-addicts
/ but for the fact that they are together [...]
/ That [**[not good]]** $_{Pred}$ **as thing**]

- ✓ NO domain delimitation: N is the generic **hyperonym** of the referent to which the predication refers (*person, thing*).
- ✓ If we remove [*come* N_{bare}] in the examples, it **would not determine an extension of the domain of predication**: in (7) he *is just funny as a person* is equivalent to *he is just funny*.
- ✓ The emptying of the semantic contribution of N_{bare} results in the construction becoming **a reprise of an already introduced referent.**

Topical-anaphoric COME

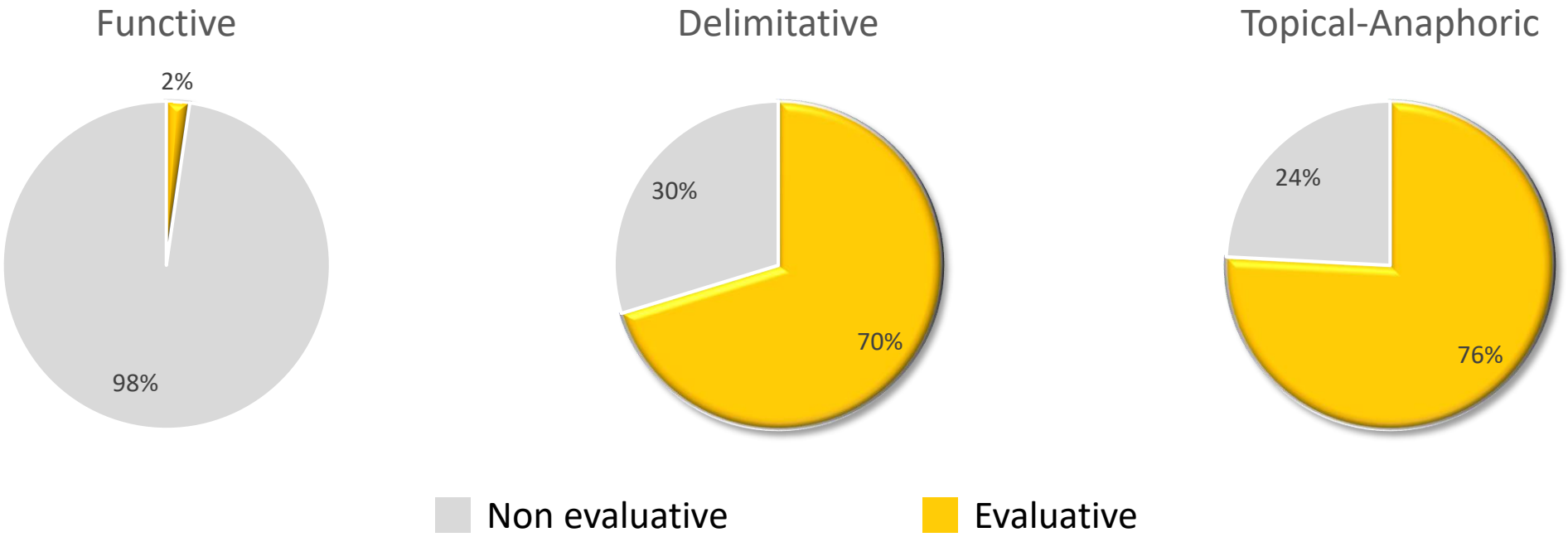
(9) perché è una città in cui stai bene / because it's a city you're comfortable in
e quindi sì abiterei a Pesaro / and so yes I would live in Pesaro
[**[mi piace]$_{Pred}$ come città**] / [**[I like it]$_{Pred}$ as a city**]
(KIParla-KIP BOD2018)

- N **resumes or rephrases the topic** already introduced

- ✓ TYPES OF N: very **rarely** animate (9%)
- ✓ PREDICATE TYPES: tendency to be **evaluative** (76%)
- ✓ CONTEXT: **anaphoric antecedent** that is often accessible but not explicit
- ✓ RARELY PREPOSED to predication (11%, es. *si' pero' comunque **[come zona]** [e' bella]$_{Pred}$*)

Evaluative contexts

- correlation between the types of functions and the **evaluative** contexts, where the speaker makes a judgment or evaluation with respect to something



The result is significant at p < .01***

Evaluative contexts and position of [*come* N].

- ✓ [P_{red} [*come* N]]
- ✓ In evaluative contexts

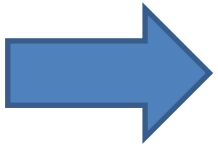
72% topical function
25% delimiting function
3% functive function

In dialogic (cf. Traugott 2010) and argumentative contexts, in which the speaker expresses an **evaluation**, we observe a further stage of 'constructionalization' (Traugott 2003), characterized by:

a. the **loss of the restrictive component** on the domain of predication and
b. the development of an **(anti-)topical function** related to syntactic dislocation (Lambrecht 1981).

Evaluative contexts and position of [*come* N]: right dislocation

- As noted by Gossen (cf. Rossi 1991), **dislocations** are most often found in contexts (such as interrogative and exclamatory sentences) in which the locutors' **emotional involvement** is strong, as is the need to emphasize the new element.
- As noted by Rossi (1991: 11), **right dislocations** are connected "**with a high degree of dialogicity**" (see also Berruto 1986).



The expression of an evaluation is a communicative act with high emotional involvement

Evaluation and dislocation of… what??

(10).	BO157:	mh	/ Mh
		dio	/ god
	BO155:	vabbe' allora ho fatto bene a non andare	/ well then I was right not to go
		non sto perdendo	/ I'm not losing
	BO157:	beh oddio insomma dai mh e'	/ well oh god i mean come on mh it's
	BO155:	il prossimo anno forse	/ next year maybe
	BO157:	**[[e' comoda]$_{Pred}$ eh come cosa]**	/ **[[it's convenient]$_{Pred}$ eh as a thing]**

(KIParla – KIP, BOA3021)

[P$_{red}$ [come cosa]]

Evaluation, dislocation and relevance

[P_{red} [come cosa]]

(10).	BO157:	mh	/ Mh
		dio	/ god
	BO155:	vabbe' allora ho fatto bene a non andare	/ well then I was right not to go
		non sto perdendo	/ I'm not losing
	BO157:	beh oddio insomma dai mh e'	/ well oh god i mean come on mh it's
	BO155:	il prossimo anno forse	/ next year maybe
	BO157:	**[[e' comoda]$_{Pred}$ eh come cosa]**	/ **[[it's convenient]$_{Pred}$ eh as a thing]**

(KIParla – KIP, BOA3021)

- ✓ Emphasis on the fact that what the speaker is stating should be related back to a certain topic that was being talked about earlier → **general anaphoric function.**
- ➢ The message relates back to what was already being talked about, therefore it is **RELEVANT**.
- ✓ Resuming the topic about which the speaker makes an assessment or evaluation increases the **relevance** of the message.

Reinforcing relevance

[P_{red} [come cosa]]

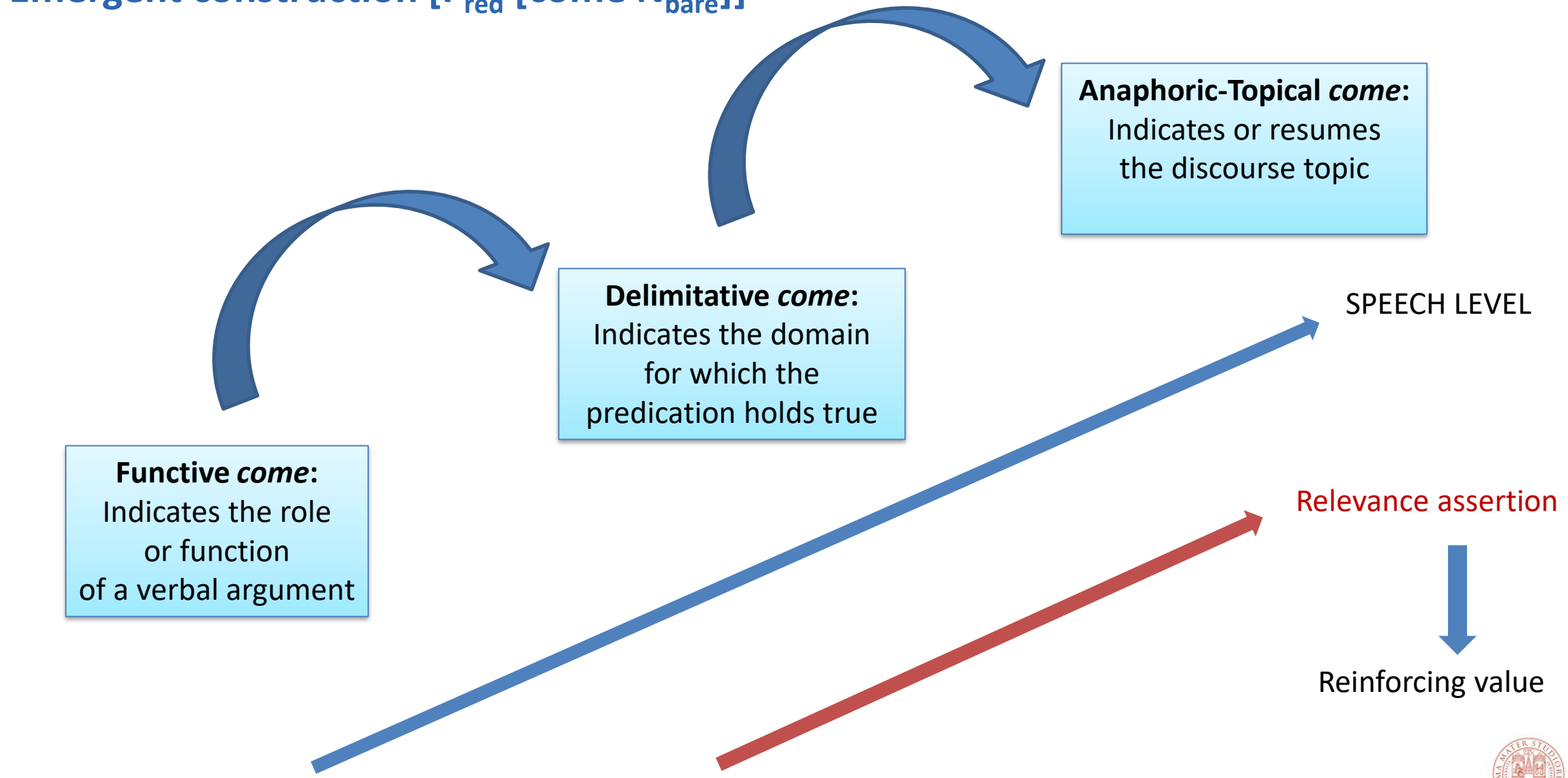
Emphasizing the **relevance of** one's assessment in conversation has the effect of **reinforcing the assessment itself**: if it is relevant, it makes sense to express it!

(11) BO156: […] il fatto che stanno insieme / the fact that they are together
BO157: che stanno insieme / that they are together
BO156: che che [**[non va bene]$_{Pred}$ come cosa**] / that [**[is not good]$_{Pred}$ as a thing**]
(KIParla-KIP BOA3021

- Shift from a function of topic delimitation > to a **reinforcing** function (cf. Voghera 2019), which emerges in dialogic and argumentative contexts

  - ✓ the speaker expresses a mostly evaluative position, and thus needs new strategies to **reinforce the relevance of his/her statement and**, consequently, his/her own **illocutionary force.**

Emergent construction [P_{red} [come N_{bare}]]



Summarizing

The construction **[P_{red} [*come* N_{bare}]]** develops functions of topic resuming to meet the need to **keep track of referents incrementally,** even in highly dialogic (and therefore potentially chaotic) interactions, such as those we find in spontaneous spoken interactions.

- Through this construction, the speaker can reestablish and reiterate as topical an accessible referent in the discourse → **discourse cohesion**

- In certain cases, the construction **[P_{red} [*come* $N_{hypernym}$]]** is used to indicate a generic connection to the preceding context → **(re)assertion of relevance**

- The expression of relevance can be instrumental in **reinforcing** assertive authority by providing a strong justification for the linguistic act - especially in dialogic, argumentative, and **evaluative** contexts

[P_{red} [come cosa]]

Overview

1. Background

2. Corpus design and implementation: goals, problems and solutions

- Modularity and incrementality
- Corpus design
- Data collection: privacy, registration and management
- Transcription of data
- Data publishing: NoSketch Engine

3. Using the corpus

- *Come cosa*: the emergence of a new construction

4. Next steps and future challenges

Next steps and future challenges

Computational developments:

- ✓ Semi-automatic transcription
- ✓ Lemmatization
- ✓ POS tagging
- ✓ UD Treebank of Spoken Italian, based on KIParla

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]. Turin: 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.

Corpus expansion:

- ✓ New modules >> *there is no data like more data*!
- ✓ Multiple levels of accessibility: ortographic, prosodically annotated, lemmatized, POS tagged, parsed.

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

www.unibo.it