Popular Italian and learner varieties: A comparative analysis of existential constructions

Silvia Ballarè (Università di Bologna)

Paolo Della Putta (Università di Torino)

Outline

1. Introduction
   - *Italiano popolare* and learner varieties
   - Existential constructions in Italian
   - Research questions

2. Data and Methods
   - KIParla corpus
   - Data extraction and annotation
   - Statistical analysis

3. Data Analysis
   - Explaining variation
   - Focus on LV

4. Conclusive remarks

Outline

Popular Italian

- Popular Italian (PI) is a social variety of Italian spoken by speakers with low educational achievements.

- It has been firstly described during the Seventies (see Cortelazzo 1972 *inter al.*), when the lower strata of the population had a diglossic repertoire.

→ Italian was used only in formal contexts and it was not their mothertongue. In informal contexts, speakers used their Italo-romance dialect.

- For this reason, PI has been described by several authors (Berruto 1983) as a **fossilized learner variety (LV) of Italian**.

- Like LV (Saldana et al. 2021; Chini 2005: 108; Guerini 2024), PI is characterized by variants resulting from acquisitional strategies, such as **paradigmatic simplification**.

Popular Italian and learner varieties

- Currently, PI is considered obsolescent due to the increased literacy levels of Italian speakers (Cerruti 2018)

- Furthermore, the sociolinguistic repertoire has changed over time (diglossia → dilalìa), with Italian becoming the mothertongue of most of the population.

- However, traces of PI remain, particularly (but not only!) among older speakers (Berruto 2012, §3; Guerini 2024).

- Conversely, immigrants who have been in Italy for many years often speak fossilized LV at basic stages of acquisition.

Popular Italian and learner varieties

- From a European functionalist perspective, interlanguages are **"continua of learner varieties"** (Klein & Perdue 1997) that, through migration, have entered the architecture of contemporary Italian (Berruto 1987; D'Agostino 2012).

- More recent works (Sobrero & Miglietta 2006; D'Achille 2016; Vedovelli 2016; D'Aguanno et al. 2019), within the SLA/Functionalist framework, highlight parallels between LV and PI (e.g. verbal paradigmatic simplification).

- Yet PI is a stabilized variety, while LVs are assumed to show diachronic mobility.

- However, studies on L2 Italian (Mocciaro 2020; Pona et al. 2024; Milano 2020) reveal strong fossilization: basic LVs persist in the contemporary Italian varietistic system.

Existential constructions in (Standard) Italian

Standard Italian

C'		*è*	*nebbia*
there		be.3**SG**	fog.**SG**

 'There is fog'

Ci		*sono*		*bambini*
there		be.3**PL**		kid.**PL**

 'There are kids'

✓ Number agreement (subject/verb)
✓ Post-verbal subject
✓ Obligatory subject

As happens in many languages of the world, in some productions a **substandard variant** occurs, with the verb appearing in singular morphology even when the subject is plural.

→ **Paradigmatic simplification**

V	S
SG	SG
PL	PL

V	S
SG	SG
	PL

See Berruto & Cerruti 2015 for an overview

Furthemore, in some varieties (centre/south) the constructions is realized with *stare* ('to stay') instead of *essere* ('to be').

Existential constructions in L2 Italian

Overgeneralization of *c'è* (SG) in early interlanguage varieties:

a) Plural subjects (Andorno & Della Putta, in prep.)

C'è molti restauranti qui in Torino 'There is many restaurants in Turin'

b) Possessive function (Andorno 2009: 13)

Io c'è lavoro buono qui Pavia 'I have (lit. I there is) a good job here in Pavia'

→ Both uses reflect general SLA tendencies (Duff 1993; Park 2021) and, in the case of b), also typological patterns (Clark 1978)

Research questions

This study aims to compare PI and LV to determine whether they exhibit similarities, focusing specifically on the behavior of a single linguistic variable: verb agreement in existential constructions (EC, Chambers 2004, Gelso 2009).

- **Are there differences or similarities in the behavior across native varieties (spoken by socially diverse speakers) and acquisitional varieties?**

- **Are there linguistic and extralinguistic factors that correlate with the lack of agreement?**

Outline

The KIParla corpus

KIParla (Mauri et al. 2019, www.kiparla.it) is a corpus of spoken Italian: it consists of 4 separate modules and it provides a large set of metadata.

- Interactions: lessons, office hours, exams, semi-structured interviews, spontaneous conversations, kitchen table conversations.

- Speakers: age range, gender, educational achievements, occupation, geographic origin.

Tokens: 2.328.193

The KIParla corpus (Stra-parla)

Currently, we are working on the creation of other two modules (i.e. Stra-ParlaBO and Stra-ParlaTO) involving speakers with a migratory background.

- Interactions: semi-structured interviews, spontaneous conversations, kitchen table conversations.

- Speakers: age range, gender, educational achievements, occupation, community, generation, age of arrival.

≈ 50 hours

Data extraction

We extracted all the occurrences of *c'è/ci sono* ('to be') and *ci sta/ci stanno* ('to stay') and we rouled out:

- Cases in which there was a singular subject;

  → sociolinguistic variable

- Cases in wich the verbs (i.e. *essere* 'to be' and *stare* 'to stay') were not involved in the scrutinized construction.

→ **Exploratory research!**

Annotation

All the occurrences were manually coded:

- Verb morphology (dependent variable):
  - sg/pl;
- NP structure:
  - noun, modifier, list, modifier + list;

SG: *c'è, ci sta*
PL: *ci sono, ci stanno*

i   ragazzi
DEF:PL.M kid:PL.M
'The kids'

i   ragazzi   italiani
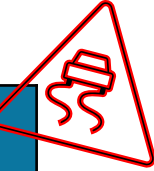DEF:PL.M kid:PL.M   italian:PL.M
'The italian kids'

il   ragazzo   e   il   cane
DEF:SG.M kid:SG.M   and   DEF:SG.M dog:SG.M
'The kid and the dog'

Annotation

All the occurrences were manually coded:

- Verb morphology (dependent variable):
  - sg/pl;
- NP structure:
  - noun, modifier, list, modifier + list;
- NP/V adjacency:
  - adjacent, non-adjacent;
- Lemma:
  - essere, stare.

Dataset

KIParla	Stra-ParlaBO/Stra-ParlaTO
N = 1.068	N = 286
Linguistic parameter	
Verb morphology, NP structure, NP/V adjacency, lemma.	
Extra-linguistic parameter	
• Register: low (spontaneous conversation, kitchen table conversation), medium (semi-structured interviews). • Educational achievements: low (middle/elementary school), medium (high school), high (university). • Age range: young (18-30), adult (31-60), senior (over 61).	
• Geographic origin: north, centre, south and islands.	• Occupation: retailer, intellectual, other. • Community: Albania, Bangladesh, China, Morocco, Ukraine. • Generation: I, II. • Age of arrival: born here, under 11, between 12 and 20, over 21.

Statistical tools

Tool	Strenghts	Limitations	Main objectives
Inference tree	• Easy to interpret; • works with unbalanced dataset.	• Overfitting; • Results may vary if the dataset is too small.	• Explore patterns; • Display distributions.
Random forest	• More robust; • works unbalanced dataset.	• Less easy to interpret, • Results may vary if the dataset is too small.	• Confirm the (relative) importance of variables → Exploratory value.
Logistic regression	• Interpretable numerical coefficients; • Can consider random effects.	• Linear relationships; • Sensitive if variables are correlated.	• Quantifies the effect of predictors.

(Gries 2020, Levshina 2015, Tagliamonte & Baayen 2012)
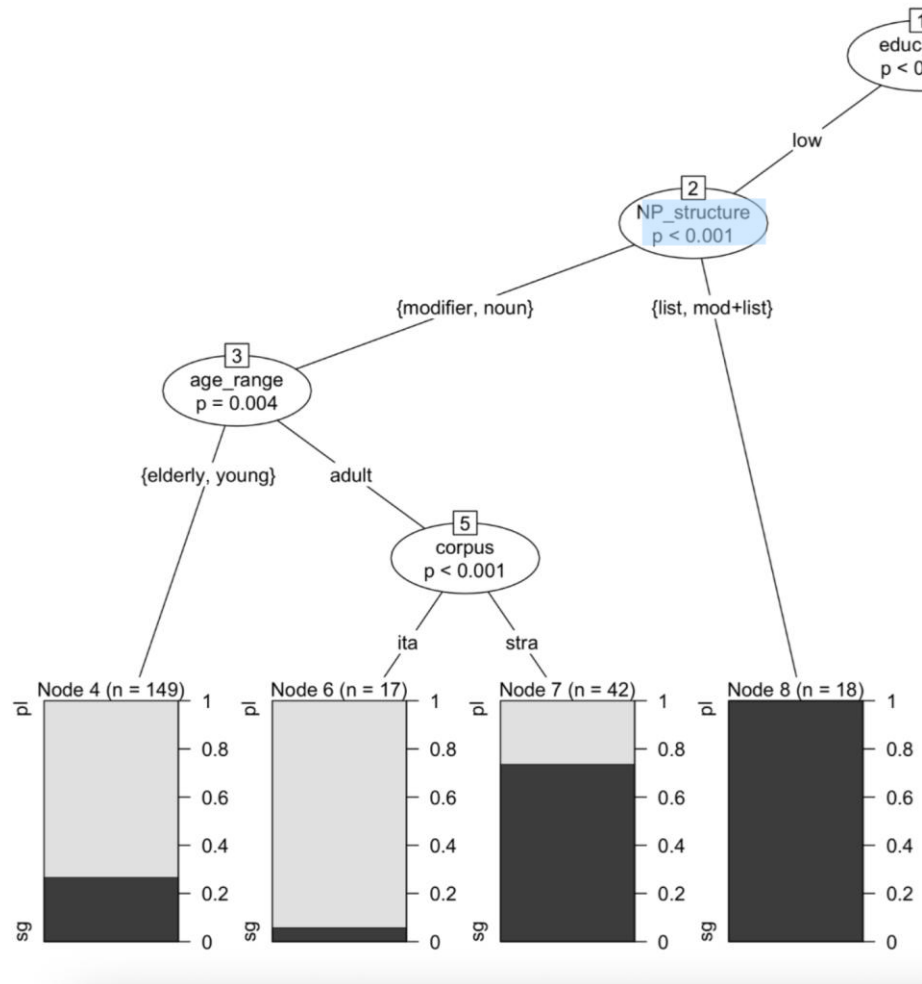
Outline

Conditional inference tree

The first parameter that is of interest in creating homogeneous sub-part of the dataset is the educational achievement of the speaker.

LOW vs. HIGH/MIDDLE

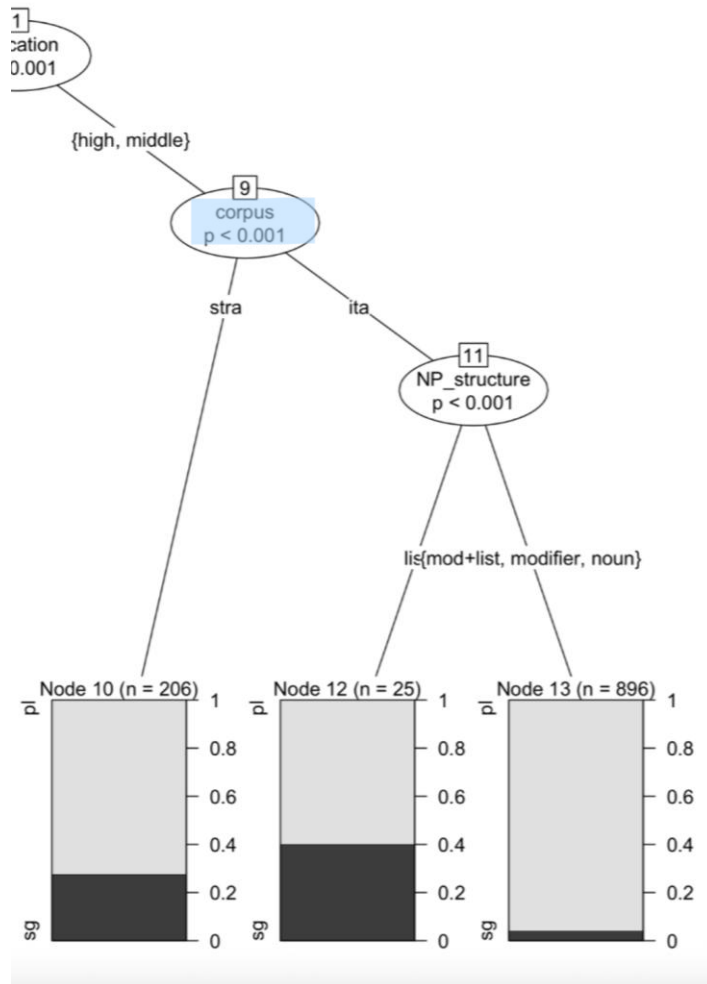C index = 0.81 → excellent discrimination

Conditional inference tree

If we consider the left side of the tree (edu = LOW), the second parameter that acquires value is the NP structure:

- List, mod+list → sg

If NP has another structure, extralinguistic parameters have importance.

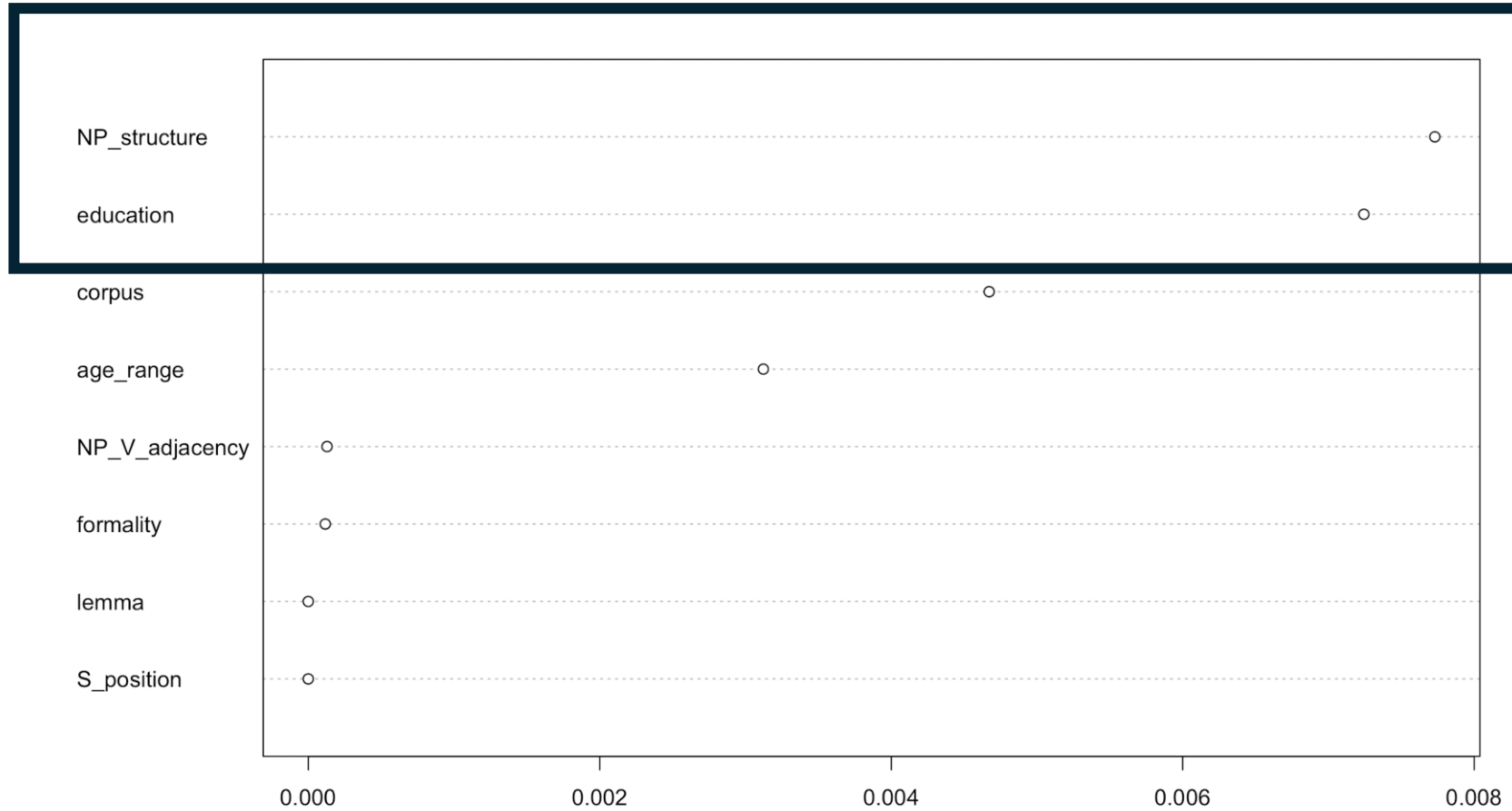C index = 0.81 → excellent discrimination

Conditional inference tree

If we move on the right side (edu = MIDDLE, HIGH), the corpus (i.e. 'italian' vs. 'speakers with migratory background') creates homogeneous subgroups.
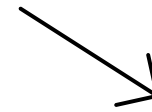
Then, only for 'italian' the NP structures acquires value.

C index = 0.81 → excellent discrimination

Random forest

The figure shows that two parameters (i.e. NP structure and education) have importance in explaining variation.

C index = 0.89 → excellent discrimination

Distributions

- **NP structure**

	modifier	noun	mod+list	list
Ci sono (pl)	85,10%	77,56%	50,00%	40,63%
C'è (sg)	14,90%	22,44%	50,00%	59,38%

- **Education**

	high	middle	low
Ci sono (pl)	88,95%	74,67%	60,18%
C'è (sg)	11,05%	25,33%	39,82%

Binomial logistic regression

Same dependent and independent variables in oreder to

- Confirm the previous (exploratory)results;

- Estimates the independent contribution of linguistic and social predictors;

- Interpet the effect size thanks to odds ratios.

Main results

Significant predictors

Significant predictors of the **generalization of the non-standard variant** (i.e. *c'è*):

- **Education (low)** → OR ≈ 7.0, p < .001 → 7 times more likely to generalize **c'è**

NP structure

- NP with modifier:

→ Strong preference for *ci sono* (OR ≈ 0.08, p < .001)

- Bare noun NP:

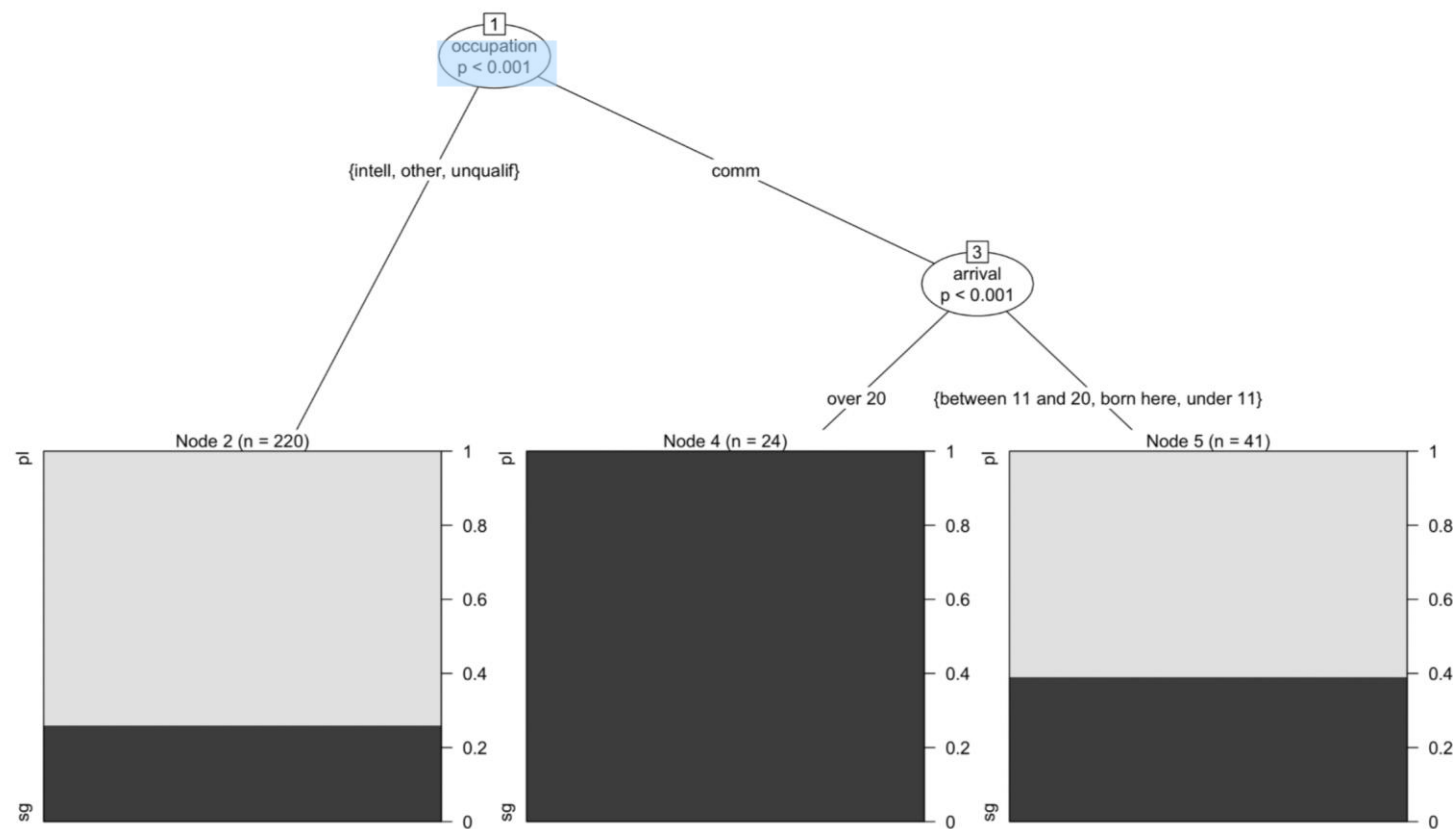→ Strong preference for *ci sono* (OR ≈ 0.12, p < .001)

- NP as list:

→ OR ≈ 0.47, p = .223 (not significant)

→ Trend towards *ci sono*, but with high variability

i   ragazzi   italiani
DEF:PL.M   kid:PL.M   italian:PL.M
'The italian kids'

i   ragazzi
DEF:PL.M   kid:PL.M
'The kids'

il   ragazzo   e   il   cane
DEF:SG.M   kid:SG.M   and   DEF:SG.M dog:SG.M
'The kid and the dog'

Conditional inference tree - LV

In this case, the first relevant parameter is the speaker's occupation:

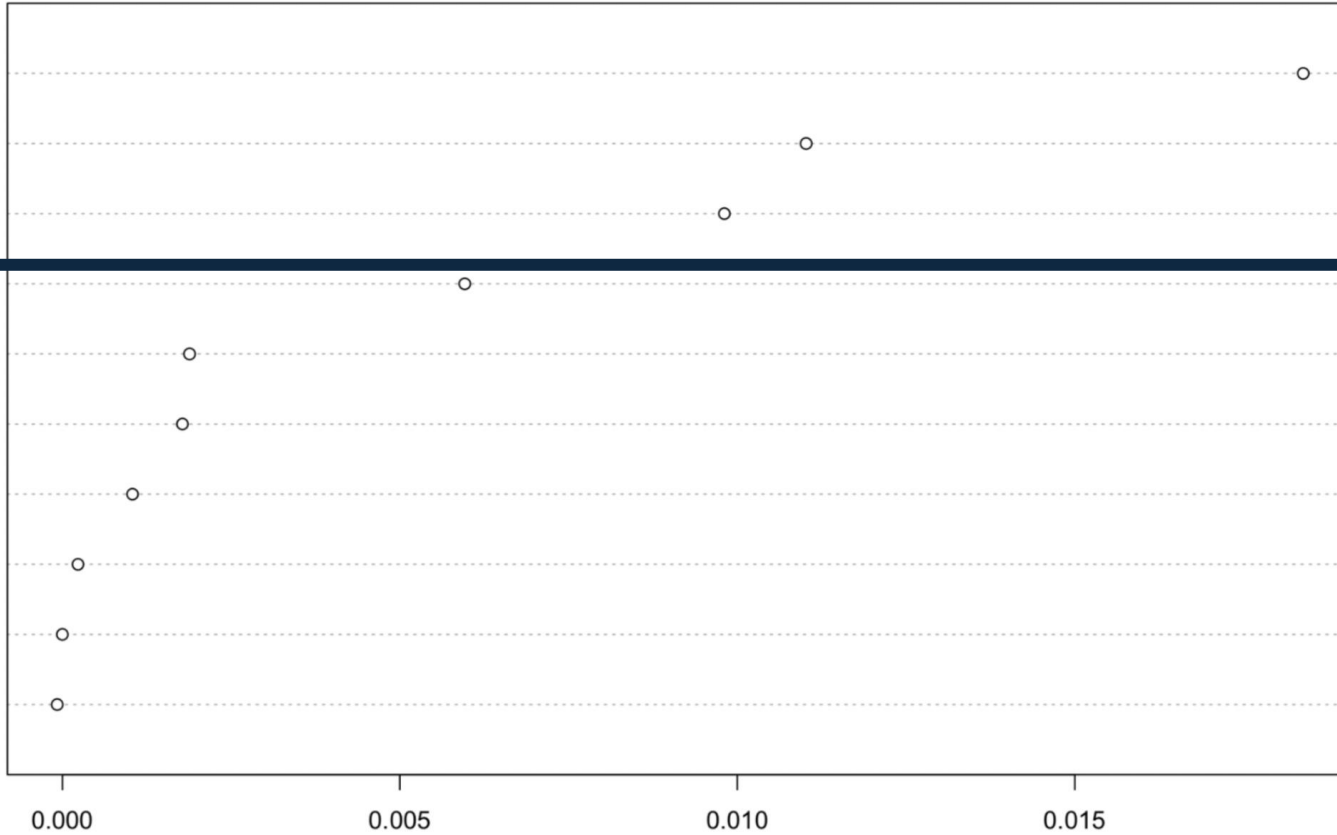Intell, other, unqualif vs. comm ('retailers')

Then, only for retailers, the age of arrival is important.

C index = 0.66 → acceptabile discrimination

Random forest - LV

LV – Distributions

Values are arranged from left to right, starting with those showing number agreement most frequently, followed by decreasing values.

- **Education**

	high	middle	low
Ci sono (pl)	70,37%	74,07%	49,40%
C'è (sg)	29,63%	25,93%	50,60%

- **Community**

	Albania	Ukraine	Morocco	China	Bangladesh
Ci sono (pl)	70,37%	74,19%	73,02%	67,06%	52,27%
C'è (sg)	0,00%	25,81%	26,98%	32,94%	47,73%

- **Occupation**

	intell	unqualif	other	retailers
Ci sono (pl)	78,26%	71,59%	60,71%	38,24%
C'è (sg)	21,74%	28,41%	39,29%	61,76%

Outline

Conclusive remarks

The analysis showed that the patterns of paradigmatic simplification of the existential construction in Italian are linked to both **linguistic** and **extralinguistic** parameters.

- As already noted in the literature, the structure of the NP plays a significant role: the singular form (*c'è*) is more frequently triggered when the subject consists of a list of coordinated elements, while it is less frequent when a bare noun is involved.

Conclusive remarks

What is most relevant from a **sociolinguistic perspective** is that speakers tend to select the substandard variant more frequently in relation to their level of education, and not according to whether they are native speakers or have a migratory background.

Speakers with a high school diploma or a university degree behave similarly, selecting the standard variant in the vast majority of cases.

Speakers with only lower secondary or primary education, on the other hand, use the substandard variant in about half of the occurrences.

For speakers with a migratory background, their community of origins and their social network seem to acquire value.

Conclusive remarks

Today, speakers of PI (i.e., with a low level of education) are very often native speakers of Italian, but they use a variety that, at least in some parts of the system, is simplified and resembles LV.

Educational qualifications do not necessarily correspond to years of schooling in Italy (→'years of instructions').

Can years of instruction be equated with the broader notion of **literacy**?

Conclusive remarks

Literacy is **"the process of using reading, writing, and oral language to extract, construct, integrate, and critique meaning in socially situated practices"** (Franken et al. 2016).

→ Low-literacy learners progress more slowly, rely on bare or periphrastic forms, are less sensitive to prosody, and avoid inflectional load (Tarone et al. 2006; De Meo & Maffia 2015; Mocciaro in press).

In LV, literacy strongly impacts L2 acquisition and may have a similar impact on native speakers.

Thank you!
Questions?

Silvia Ballarè (Università di Bologna)

Paolo Della Putta (Università di Torino)

Language and languages at the crossoreads of disciplines
Lille, 01-03/09/2025