The Longitudinal Corpus of Chinese Learners of Italian (LoCCLI)

Università per Stranieri di Perugia

Stefania Spina - *University for Foreigners of Perugia, Italy*
Anna Siyanova-Chanturia - *Victoria University of Wellington, New Zealand*

Te Whare Wānanga o te Ūpoko o te Ika a Māui Victoria University of Wellington

1 Introduction

Learner Corpus Research on Italian is fairly recent, compared to other languages: the first learner corpora of Italian were released in 2009.

Features of existing learner corpora of Italian:
- small size
- mainly cross-sectional
- data often collected opportunistically, without systematic design criteria.

Motivation:

- Need for longitudinal learner data of Italian, collected by following accurate and systematic design criteria.

The ***Longitudinal Corpus of Chinese Learners of Italian*** (**LoCCLI**) is the first large-scale longitudinal corpus of Italian as a second language.
It was started in 2016 and is available via CQPweb (https://www.unistrapg.it/cqpweb/).

2 Method

Participants:
- 175 Chinese learners, age 17-33 (mean=20.5, SD=2.7; 105 females).

Time spent in Italy:
- On average, 1.7 months (range 0.5-5, SD=0.69) before writing the first essay.

Collection:
- 2 data collection points: at the beginning of a six-month course of Italian, and at the end of the course.

Task:
- Each of the 175 learners contributed two written essays on two of the following topics (350 total essays):
  1) *My first impression of Italy and Italians*
  2) *My hobbies: what do I usually do in my free time*
  3) *My last holidays.*

Proficiency level:
- Through a placement test, learners were assigned to one of three proficiency levels: A1 (n=39), A2 (n=86), and B1 (n=50).

Annotation:
- pos-tagging and lemmatization (using an ad hoc version of TreeTagger)
- xml annotation.

Size:
- 97,000 tokens

	A1	A2	B1
data collection a	7,126	22,851	15,903
data collection b	9,487	24,117	17,386

Number of tokens for data collection point and proficiency level

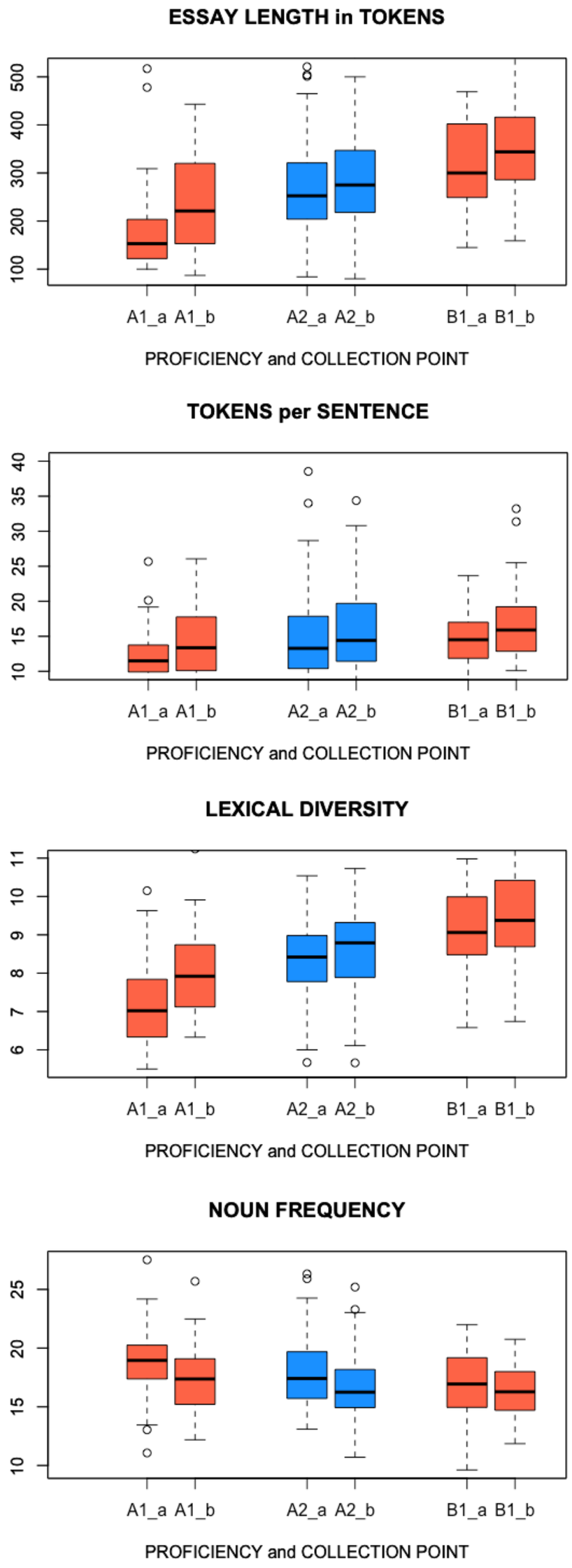
4 Papers based on the LoCCLI

Siyanova-Chanturia & Spina (in preparation). *Longitudinal investigation of multi-word expression development in learner writing*.
Spina (in preparation). *The development of phraseological errors in Chinese learner Italian: a longitudinal study*.
Spina (forthcoming). *Lo sviluppo longitudinale della fraseologia in apprendenti cinesi di italiano L2: uno studio preliminare su alcune categorie di errori*. In Ricognizioni.
Forti (2017). *Data-driven learning and the acquisition of Italian collocations: from design to student evaluation*. In K. Borthwick, L. Bradley & S. Thouësny (Eds), CALL in a climate of change: adapting to turbulent global conditions – short papers from EUROCALL 2017.

Further information

Stefania Spina (stefania.spina@unistrapg.it)
Anna Siyanova-Chanturia (anna.siyanova@vuw.ac.nz)

3 Descriptive statistics on vocabulary development



Texts collected after six months **differ in terms of lexical diversity** and learners' ability to produce **longer sentences and texts**, rather than in terms of different distribution of grammatical categories (with the exception of nouns). This is particularly clear in **beginner** (A1) and **intermediate** learners (B1).

5 Further work

- investigation of multi-word expression development in Chinese learners of Italian;
- creation of a native counterpart of the corpus (the LoCCLI-IT);
- dependency parsing of the LoCCLI, in order to gain accuracy in multi-word expression extraction.