Received: 8 March 2022 | Deposited: 25 March 2022 | Published: 5 September 2022

DOI: 10.7302/66g9-e028

DATASET

The Lannang Corpus (LanCorp):

A POS-tagged, sociolinguistic corpus containing recordings and transcriptions of Lannang speech collected from the metropolitan Manila Lannangs between 2016 and 2020

Wilkinson Daniel Wong GONZALES[1]

[1]Department of English
The Chinese University of Hong Kong
Hong Kong SAR, People's Republic of China

Correspondence
Wilkinson Daniel Wong Gonzales, Department of English, The Chinese University of Hong Kong, 321 Fung King Hey Building, Hong Kong SAR, People's Republic of China 999077
Email:
wdwonggonzales@cuhk.edu.hk;
wdwg@umich.edu

Funding information
The University of Michigan Department of Linguistics, Lieberthal-Rogel Center for Chinese Studies, International Institute

Overview

The Lannang Corpus (LanCorp) is a sociolinguistic POS-tagged 375,000-word speech-and-text corpus of Lannang languages based on audio recordings collected in metropolitan Manila between 2016 and 2020. It hopes to furnish scholars interested in Sino-Philippine (socio)linguistics with a contemporary, multilingual corpus (i.e., Hokkien, Tagalog, English, Lánnang-uè, Mandarin) compiled using recorded oral data primarily collected from a Sino-Philippine community in metropolitan Manila by the community: the Manila Lannangs. The publicly available corpus contains manual transcriptions (time-aligned to the audio), source language and part-of-speech tags derived using a mix of manual and computational methods, and a wide range of social metadata; it is also organized and stored systematically for easy data retrieval and (socio)linguistic analysis. Although there are existing sociolinguistic corpora, they are small in scale and were not released publicly due to lack of informant consent – LanCorp readily fills the gap.

Methodology

The data are output from sessions conducted with 119 "Lannang" community members in metropolitan Manila (Gonzales 2021; Gonzales 2022a; Gonzales 2022b; Gonzales 2022c). In the sessions, participants were instructed to tell a story using the wordless book *Frog, Where Are You?*

By Mercer Mayer in Lánnang-uè but were also encouraged to use other languages that they feel comfortable using (e.g., Filipino) (Gonzales 2022c). The participants were then interviewed using a set of questions that focus on questions about community, identity, language, and education. Although the language used by the interviewer is Lánnang-uè, participants were not restricted to speak in a particular language. Participants provided several sociolinguistic information through a survey.

Data in unstructured casual conversations was collected differently. A Lannang data collector reached out to five Lannang families and asked for verbal consent to record their conversations in their gatherings (e.g., restaurants, home) (Gonzales 2016). A microphone was then placed in an inconspicuous location in these gatherings for roughly thirty minutes. The data was collected in 2016 and submitted to the corpus compiler 2020; it was not linked to social metadata upon the request of the participants, who also requested that the raw audio file not be released.

A total of 22 trained individuals – all fluent in Lannang languages – were trained to use ELAN (The Language Archive 2020) and were familiarized with Lannang Orthography conventions (The Lannang Archives 2020). Upon passing the summative transcription assessment, they were assigned a portion of the audio files and were instructed to use ELAN to (1) segment the audio files into sentences, (2) identify the source language of the sentence, and (3) transcribe each sentence into text using Lannang Orthography conventions (The Lannang Archives 2020). I instructed four of these transcribers to go over the files and transcriptions. All participants were paid for their time.

For POS-tagging and source-language tagging, the ELAN files were first converted to spreadsheet files. From these, I extracted all Lánnang-uè sentences. Then, I tagged each word in all sentences for part-of-speech (POS) (e.g., conjunction, preposition) using a POS-tagger program I created in the Python environment (Van Rossum and Drake 2009). The program utilizes Conditional Random Fields (CRF) (Lafferty et al. 2001) – a model that 'learns' the POS distributions from sequential data and can identify the optimal POS of a token, given the context. The CRF model used was trained using 1,085 manually annotated Lánnang-uè sentences. It has a cross-validated (k-folds = 5) accuracy score of 0.83 (SD = 0.005), precision score of 0.58 (SD = 0.017), recall score of 0.56 (0.018), and an f-1 score of 0.56 (SD = 0.015).

After POS tagging, I tokenized the sentences – I broke down the tagged sentences into tagged words. I then tagged each word for source language by relying on a combination of rule-based and manual tagging approaches. I used publicly available English, Tagalog, and Mandarin wordlists to help me tag English-, Tagalog-, and Mandarin-origin Lánnang-uè words. Lánnang-uè words that are not found in any of the three wordlists are preliminarily tagged as Hokkien-sourced. I hired and trained three native speakers of Lánnang-uè to go over the list and revise incorrectly tagged tokens. I also asked them to tag words that do not have a clear origin as 'unclear'.

The resulting LanCorp is not only transcribed, time-aligned, and searchable, it is also tagged with source language information and part-of-speech.

Instrument and/or software specifications

A Zoom H6 recorder was used to record the participants. The recordings were then analyzed using ELAN (The Language Archive 2020).

Files contained in the submission

The LanCorp folder contains four subfolders: (1) Corpus in audio and ELAN format, (2) Corpus in text format, (3) Corpus in spreadsheet format, and (4) Sociolinguistic metadata. The first folder contains two more subfolders, each respectively containing WAV files and .eaf files. The second folder contains .txt files. The third and fourth folders are .csv files (the sociolinguistic metadata can be merged with the .csv corpus using R or other data processing software).

The audio and ELAN files are labelled as follows: uniqueuserid-contextYEAR", where context refers to the stylistic condition in which the recording was done (i.e., CLIN = interviews, FRST = frog story narrative, PROT = unstructured conversations) and YEAR refers to the last two digits of the year of recording.

A similar convention is followed for each utterance line in the spreadsheets: each utterance is tagged with a <style-year-uniqueuserid-uniqueutteranceid> metadata tag. For instance, the tag <CLIN-18-68:1> indicates that the utterance was recorded as part of an interview conducted in 2018. It indicates that the utterance was produced by an individual with the identification number 68 and that the utterance has a unique identification number of 1.

For instance, the label CLIN-18-68 indicates that the utterance was recorded as part of an interview conducted in 2018. It indicates that the utterance was produced by an individual with the identification number 68 and that the utterance has a unique identification number of 1.

Metadata

Each utterance in LanCorp is linked to file-, utterance-, and participant-level metadata, which can be used for (socio)linguistic analyses (Gonzales 2018; Gonzales 2022c). There are 13 file- and utterance-level variables and 40 participant-level sociolinguistic variable (groups). The complete list of available metadata is provided in TABLE 1 and TABLE 2, where I also note the type of metadata/variable and provide examples for each of them.

TABLE 1 File- and utterance-level metadata linked to the Lannang Corpus

variable	variable name in corpus	type	example
tag	tag	categorical	<CLIN-18-68:1>
unique utterance number	no	continuous	1, 2, 3...
ELAN file	filename	categorical	PC0001-CLIN18.eaf
year collected	year	continuous	18 (2018), 19 (2019)
region collected	region	categorical	MNL (Manila)
audio file time alignment (begin/end) in minutes, seconds, milliseconds	begin, end	time	00:03.0
duration of utterance in minutes, seconds, milliseconds	duration	time	00:04.3
interlocutor identification	interlocutor.no	categorical	1, 2, 3...

language of utterance	lg.clause	categorical	l (Lánnang-uè), x (unknown), t (Tagalog), g (English with Tagalog borrowings), etc.
context of utterance	context	categorical	CLIN (interview), PROT (unstructured conversation), FRST (narrative)
transcriber identification	transcriber.id	categorical	1, 2, 3…
file index	index.file	categorical	1, 2, 3…
relative point in audio file - in percentage	filepercent	categorical	0.002298851

TABLE 2 Informant-level sociolinguistic metadata linked to the Lannang Corpus

variable	**variable name in corpus**	**type**	**example**
individual identification number	idno	continuous	1,2,3
date of collection	date.of.collection	continuous	2019, 2020
occupation	occupation	categorical	student, pastor
Chinese ancestral origin of mother (by province)/ (by city)	origin.ancestral.province.maternal	categorical	Fujian, Shishi
Chinese ancestral origin of father (by province)/ (by city)	origin.ancestral.province.paternal	categorical	Guangdong, Taishan
age	age	continuous	25
sex	sex	categorical	M, F
distance from Binondo (average)/ (place of current residence)/ (place of longest residence)	distance.average, distance.curres.binondo, distance.longres.binondo	continuous	2.7
residence in Manila (at all)/ (current)/ (longest)	bin.manila, bin.curres.manila, bin.longres.manila	categorical	1,-1
religion (protestant or not)	bin.christian	categorical	1,-1
religion	religion	categorical	Protestant, Catholic
institution of worship	institution.worship	categorical	UECP
identity as Filipino/ Filipino-Chinese/ Chinese Filipino/	identity.fil, identity.filchi, identity.chifil, identity.huakiau,	categorical	1,0

Huakiau/ Chinese immigrant/ Chinese	identity.chineseimmigrant, identity.chinese		
identity (main)	main.identity	categorical	Lannang, Chinese Filipino
generation	generation	numerical	2,3,4
language used with mother	language.with.mother	categorical	L (Lánnang-uè), H (Hokkien)
language used with father	language.with.father	categorical	T (Tagalog), CD (Taishanese)
native language is Tagalog/Philippine Hokkien/ Lánnang-uè/ English/Standard Cantonese/ Taishanese/ Standard Hokkien/ Mandarin	native.language.Tagalog, native.language.PHHokkien, native.language.Lannangue, native.language.English, native.language.StandardCantonese, native.language.Taishanese, native.language.StandardHokkien, native.language.Mandarin	categorical	1,0
first language	language.L1	categorical	L (Lánnang-uè), H (Hokkien)
second language	language.L2	categorical	T (Tagalog), CD (Taishanese)
language used at home	language.home	categorical	L (Lánnang-uè), H (Hokkien)
highest level of education	education.highest	categorical	Doctorate, Masters
place of education (elementary)/ (high school)/ (college)	school.elementary, school.highschool, school.college	categorical	Hope Christian High School
number of languages used in education (elementary)/ (high school)/ (college)	number.moi.elementary, number.moi.highschool, number.moi.college	continuous	3
medium of instruction in Chinese education (elementary)/ (high school)/ (college)	moi.elementary.chinese, moi.highschool.chinese, moi.college.chinese	categorical	L (Lánnang-uè), H (Hokkien)
place of current residence (city)/(district)	residence.current.city, residence.current.district	categorical	Manila
place of longest residence (city)/(district)	residence.longest.city, residence.longest.district	categorical	Manila
latitude/longitude of current city of residence	residence.current.city.latitude, residence.current.city.longitude	continuous	14.599512, 120.984222

latitude/longitude of longest city of residence	residence.longest.city.latitude, residence.longest.city.longitude	continuous	14.599512, 120.984222
length of stay in the Philippines in years	years.in.philippines	continuous	15, 49
percentage of life in the Philippines	length.stay.philippines.proportion	continuous	0.6
length of stay in Manila in years	years.in.manila	continuous	40, 49
percentage of life in Manila	length.stay.manila.proportion	continuous	0.2
self-reported proficiency in Tagalog/ Hokkien/ Lánnang-uè/ English/ Cantonese/ Taishanese/ Mandarin	prof.t, prof.h, prof.l, prof.e, prof.cs, prof.cd, prof.m	continuous	1 to 7
self-reported comfort in languages (see list above)	comf.t, comf.h, comf.l, comf.e, comf.cs, comf.cd, comf.m	continuous	1 to 7
self-reported confidence in languages (see list above)	conf.t, conf.h, conf.l, conf.e, conf.cs, conf.cd, conf.m	continuous	1 to 7
self-reported frequency of language use - languages (see list above)	freq.t, freq.h, freq.l, freq.e, freq.cs, freq.cd, freq.m	continuous	1 to 7
self-reported pride in languages (see list above)	pride.t, pride.h., pride.l, pride.e, pride.cs, pride.cd, pride.m	continuous	1 to 7
self-reported importance of language in society - languages (see list above)	imp.t, imp.h, imp.l, imp.e., imp.cs, imp.cd, imp.m	continuous	1 to 7
attitudes - Lánnang-uè is barok/ konyo/ comical/ natural/ bad-sounding/ bastardized/ prestigious/ reflective of the Lannang identity	barok.l, conyo.l, comical.l, natural.l, badsounding.l, bast.l, prestige.l, ref.l	continuous	1 to 7

Use and Access

This data set is made publicly available under the Attribution-Non Commercial 4.0 International (CC BY-NC 4.0) license. It can be accessed in Deep Blue Data, a repository hosted by The University of Michigan, using the DOI link provided: https://doi.org/10.7302/66g9-e028

Consent and ethics

Participants have given consent to release the linguistic and sociolinguistic data on the condition that their names not be made public, and that the data will be used only for academic/non-commercial purposes (e.g., sociolinguistic analyses, social analyses, historical analyses). The collection protocol has been vetted by the University of Michigan Institutional Review Board in 2019.

Declaration of Competing Interest

The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

GONZALES, WILKINSON DANIEL WONG. 2016. Trilingual code-switching using quantitative lenses: An exploratory study on Hokaglish. *Philippine Journal of Linguistics* 47.106–128.

GONZALES, WILKINSON DANIEL WONG. 2018. Philippine Hybrid Hokkien as a postcolonial mixed language: Evidence from nominal derivational affixation mixing. Singapore: National University of Singapore master's thesis.

GONZALES, WILKINSON DANIEL WONG. 2021. Filipino, Chinese, neither, or both? The Lannang identity and its relationship with language. *Language & Communication* 77.5–16.

GONZALES, WILKINSON DANIEL WONG. 2022a. Hybridization. *Philippine English: Development, Structure, and Sociology of English in the Philippines*, ed. by Ariane Macalinga Borlongan, 170–183. London: Routledge.

GONZALES, WILKINSON DANIEL WONG. 2022b. Interactions of Sinitic Languages in the Philippines: Sinicization, Filipinization, and Sino-Philippine Language Creation. *The Palgrave Handbook of Chinese Language Studies*, ed. by Zhengdao Ye, 369–408. Singapore: Springer Nature Singapore. doi:10.1007/978-981-16-0924-4_31. https://link.springer.com/10.1007/978-981-16-0924-4_31.

GONZALES, WILKINSON DANIEL WONG. 2022c. "Truly a Language of Our Own" A Corpus-Based, Experimental, and Variationist Account of Lánnang-uè in Manila. Ann Arbor: University of Michigan ph.d. dissertation.

LAFFERTY, JOHN.; ANDREW MCCALLUM.; and FERNANDO C. N. PEREIRA. 2001. Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the 18th International Conference on Machine Learning*.282–289.

THE LANGUAGE ARCHIVE 2020. ELAN. Nijmegen, The Netherlands: Max Planck Institute for Psycholinguistics, The Language Archive. https://archive.mpi.nl/tla/elan.

THE LANNANG ARCHIVES 2020. Lannang Orthography. *The Lannang Archives*. The Lannang Archives. https://www.lannangarchives.org.

VAN ROSSUM, GUIDO.; and FRED L. DRAKE. 2009. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

To Cite the Dataset

Gonzales, Wilkinson Daniel Wong. 2022. The Lannang Corpus (LanCorp): A POS-tagged, sociolinguistic corpus containing recordings and transcriptions of Lannang speech collected from the metropolitan Manila Lannangs between 2016 to 2020. Deep Blue Data, Deep Blue Repositories. The University of Michigan Library. DOI: https://doi.org/10.7302/66g9-e028